Conversation-Based Assessments:

Measuring Student Learning with Human-Like Communication

by

Seyma Nur Yildirim Erbasli

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

MEASUREMENT, EVALUATION AND DATA SCIENCE

Department of EDUCATIONAL PSYCHOLOGY

University of Alberta

Abstract

In recent years, conversational agents have been widely used in education to support student learning. Conversational agents have the capability to enhance learning by improving interaction, motivation, feedback, and personalization. To date, researchers have designed and used different types of conversational agents—such as virtual teaching assistants, tutors, and peers (or learning companions)—to support the student learning process. In addition to the instructional use of conversational agents, there have been new attempts in recent years to design and use conversational agents for educational assessments (i.e., conversation-based assessments: CBA), with the goal of improving students' assessment experiences. To address the limited research on CBA, this research aimed to design a CBA as a formative assessment to assess higher education students' knowledge and provide support and feedback to scaffold their learning. This study introduced a CBA with selected-response (multiple-choice and true-false) and constructed-response (short-answer) tests and evaluated its performance based on intent classification and confidence score. CBA was designed using Rasa—an artificial intelligence-based tool—and deployed to Google Chat to share with students in two sections of an undergraduate-level course, Educational Assessment. One section of the course (n_1 = 290) was provided with only the selected-response format while the other section (n_2 = 119) was provided with both CBA formats following course instructors' availability and preference to use CBA in their sections. A survey was administered after students experienced CBA to investigate their attitudes toward CBA. The unique total number of students who took selected-response and constructed-response tests and completed the survey are 98, 21, and 61 respectively. In addition, CBA was evaluated by students in another undergraduate-level course, Introduction to Human Computer Interaction, (n_3 = 106) to find usability issues through a cognitive walkthrough. The conversation data showed

that CBA with both selected-response and constructed-response items produced high standard accuracy measures and confidence scores for each intent (i.e., student response). CBA with selected-response items interpreted all student responses accurately and chose the appropriate conversation paths (F1-measure of 100% and the confidence score of 1 for each intent). In comparison, CBA with constructed-response items consistently matched student responses to the appropriate conversation paths for the most part (F1-measures ranged from 89% to 100%, and confidence scores ranged from 0.30 to 0.99). The findings suggest that ensuring the accuracy of CBA with constructed-response items is more challenging than CBA with selected-response items. According to the survey data, most of the students reported positive attitudes toward CBA. Student reactions to CBA and regular assessments (e.g., online quizzes) were very similar. The findings from the cognitive walkthrough of CBA showed its usability, however, several important usability problems were also reported to improve the user interaction with CBA. Highly accurate dialogue moves within CBA, positive student attitudes toward CBA, and usability indicators suggest the utility of CBA in measuring student knowledge and skill as well as enhancing their assessment experiences. Overall, this study indicates the promise of conversational agents in developing more interactive assessments to measure higher education students' knowledge and skill as well as enhance their assessment experiences through a more interactive assessment environment.

Keywords: conversational agents, conversation-based assessment, artificial intelligence, natural language understanding

Preface

This thesis is an original work by Seyma Nur Yildirim Erbasli. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "Real-Time Assessment and Feedback for Real-Time Learning with Conversation-based Assessments", No. Pro00114191, 2021-11-23.

Some part of Chapter 2 has been published as Yildirim-Erbasli, S. N., & Bulut, O. (2021). Conversation-based assessments: Real-time assessment and feedback. *eLearn Magazine, 12,* Article 1. doi: 10.1145/3508017.3495533. Yildirim-Erbasli was responsible for the manuscript composition. Bulut assisted with the manuscript and contributed to manuscript edits.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1: Introduction

One-on-one tutoring can be extremely effective compared to a typical classroom assessment (Corbett, 2001; Fletcher, 2003) because human tutors can scaffold student learning with the help of dialogue (e.g., Chi et al., 2008). Tutors may assign tasks for students to solve, then review their responses to gain a better understanding of what they know. If the tutor suspects a misunderstanding, they may ask more questions and repeat the process with several questions and responses. The additional questions may demonstrate a misunderstanding or that the student understands the subject but was unable to deliver a correct answer initially for some other reason. This type of interaction reveals what the student understands and can do, as well as areas where additional learning is required. In comparison to a non-interactive approach, this adaptive strategy allows students to communicate their knowledge while also providing the teacher with more diagnostic information (Jackson & Zapata-Rivera, 2015). These human-to-human interactions can provide useful insight and evidence for assessment purposes, and while they are fairly easy to utilize on a small number of students, they are neither easy nor financially sustainable with a large group of students. One feasible strategy to provide learning assistance to all students is to use conversational agents.

Research into conversational agents has a long history starting as early as 1966. ELIZA (Weizenbaum, 1966) was the first reasonably successful, popular, and widely used conversational agent. It simulated a client-centered psychotherapist and turned a patient input into a therapist question through simple syntactic transformational rules (i.e., formulating rules by detection of keywords and word combinations). Since then, efforts to build conversational agents have continued. Intelligent tutoring systems (ITS) were launched in the late 1970s as computerized learning environments that optimized each student learning (D'Mello & Graesser,

2013). ITS adaptively responds and gives immediate feedback to student actions, and guides students on what to do next in a fashion that is designed to estimate what students know (Graesser et al., 2014). ITS was nearly as effective as human tutoring in helping students learn (Corbett, 2001; Dodds & Fletcher, 2004; VanLehn, 2011; Wisher & Fletcher, 2004). However, ITS considered students as passive knowledge recipients and did not bring the notion of interaction within the learning environment. The concept of interaction has been introduced to the learning environment by conversational agents.

Thanks to recent breakthroughs, there has been an increase in the use of conversational agents. Over the last decades, technical breakthroughs, and advances in the fields of computational linguistics, information retrieval, cognitive science, artificial intelligence (AI), and discourse processes have offered affordances to researchers to build successful conversation (or dialogue) systems (Graesser et al., 2014). Global technology corporations such as Microsoft, Google, and Amazon have been working on AI-based agents for decades and have made them available to the public. AI-based agents bring more interaction and intelligence to conversational agents than earlier generations of agents (Maedche et al., 2019).

Researchers have developed AI-based conversational agents in education to personalize and improve student learning (Goel & Joyner, 2017; Graesser et al., 2014). Although the majority of research on conversational agents has been done for tutoring purposes, researchers have lately begun to investigate and apply this method in the assessment field (e.g., Jackson et al., 2018). Conversation-based assessments (CBA), which take place between a student and a computer, can provide innovative and interactive assessments. The idea behind CBA is to use automated or adaptive conversations to measure and support student learning by taking an assessment with a computer agent through conversations. CBA combines assessment and

feedback in an optimal formula to improve student learning while assessing their knowledge and providing timely feedback. Thus, unlike conventional digital assessments, CBA has more potential to measure but also improve student learning and motivation. The advantages of CBA inspired this study to motivate students to take assessments by holding conversations and to support and scaffold their learning by providing timely feedback.

Theoretical Framework

Constructivist Learning Theory

Constructivism is an approach to learning that students construct knowledge actively rather than receive information passively. According to constructivism, learning is not a passive activity; it requires students to take action (Snowman & McCown, 2015). Students must be actively involved in their learning and growth through engaging in the world. In constructivism, students develop their representations and incorporate new information into their pre-existing knowledge. They, in general, use their prior knowledge, experiences, beliefs, and insights as a foundation and then build on it with new information (Snowman & McCown, 2015). According to the constructivist learning theory, learning is more effective and deeper when students actively generate answers than when they are merely given information (Graesser et al., 2014). It emphasizes the importance of social connection in learning through conversation and interaction to assist student learning. Learning is an activity that requires interaction as an ability to learn and is inextricably linked to social connections. Thus, to obtain meaningful learning, students need to actively construct knowledge by interacting with others. CBA provides this constructivist environment to students. It stimulates dialogue moves between students and human tutors (or peers) who guide students in the construction of knowledge by asking

questions, giving prompts or hints, and providing feedback and explanations (D'Mello & Graesser, 2013).

Socio-Cultural Theory

"The distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem-solving under adult guidance, or in collaboration with more capable peers" is known as the zone of proximal development (ZPD) (Vygotsky, 1978, p.86). Human tutors adapt the level of conversation to the student ZPD based on their assessment of the student competence. The scaffolding of human tutors can and should be emulated by conversational agents to support student learning. In line with the theory of ZPD, CBA can help students to achieve a level of comprehension that is closest to their ZPD with appropriate scaffolding. A conversation-based approach can scaffold student ZPD by adapting its conversations based on student initial responses. CBA can provide details about ZPD by using hints or prompts to obtain information that the student already knows but did not include in their initial response (Jackson et al., 2018). CBA may dynamically assess student knowledge state and effectively alter the conversations (learning material) in terms of both content and pedagogy to give relevant instruction and support to student learning. Thus, CBA can tailor the interaction between the computer agent and student toward the student ZPD. For example, when a student provides an incorrect answer, the conversational agent attempts to help the student recorrect their incorrect answers by generating hints and scaffolding their ZPD. In CBA, students may have a variety of options and scaffolds to assist them in producing a comprehensive response (Jackson & Zapata-Rivera, 2015).

Self-Regulation

Self-regulated learning is the ability to comprehend and control various aspects of a learning environment. Self-regulated learning occurs when students create their learning goals and then regulate and control their actions to achieve their goals (Zimmerman, 1990). It includes the following steps: set goals, plan strategies, perform strategies, monitor performance, and reflect on performance. Therefore, self-regulated learners ask questions, seek answers, and evaluate the quality of the answers to satisfy their personal curiosity. Self-regulated learning includes three major aspects of learning: cognition, metacognition, and motivation (Snowman & McCown, 2015). Cognition consists of mental processes (e.g., thinking, learning, knowing, remembering, judging and problem-solving) to gain knowledge and comprehension. Metacognition enables students to understand and monitor their cognitive processes. It helps students become aware of their strengths and weaknesses. Motivation affects the development and use of cognitive and metacognitive skills. In the context of assessment, it is hypothesized that students with a high level of motivation are expected to engage with the assessment while those with a low level of motivation are more likely to disengage (e.g., Wise & Kong, 2005). CBA can be a potentially practical solution to the poverty of self-regulation by supporting cognition, metacognition, and motivation through its interactive and personalized assessment environment.

Cognitive Development

Confusion is likely to arise during cognitive disequilibrium (Kort et al., 2001; Rozin & Cohen, 2003). Following the definition of disequilibrium by Piaget—cognitive imbalance, i.e., a mismatch between what is learned and what is encountered—students can be in cognitive disequilibrium when they are challenged with a problem or question. These challenging

questions reflect personal curiosity and thus support self-regulated learning. When students are challenged and confused, they are in a state of cognitive disequilibrium. At this point, CBA can effectively promote learning with dialogue moves to manage the confusion productively. Regarding being in cognitive disequilibrium or being challenged and confused, researchers investigated what emotions students experienced when they interacted with a conversational agent (Craig et al., 2004; D'Mello et al., 2007; Graesser et al., 2007). They reported frustration (i.e., angry or agitated), boredom (i.e., uninterested or slow response), flow (i.e., interest or attention or quick response), confusion (i.e., puzzled or not sure or struggling), eureka (i.e., transfer from a state of confusion to a state of intense interest), and neutral (i.e., void of emotion or no facial features or no emotions determined). Among these emotions, only confusion predicted student performance on a post-test (Graesser et al., 2007) and there was a positive correlation between confusion and learning (Craig et al., 2004). Moreover, students who were confused during the learning session performed better than those who were not (Craig et al., 2004).

Conversation-Based Assessments

CBA involves conversations among computer-animated agents and test-takers. They can be used to generate assessments, which can be used to provide feedback to students or teachers. CBA helps students learn by holding a conversation by adapting the conversation to student responses (Graesser, 2016). CBA may help students to behave in a manner similar to how they would experience a typical conversation on a particular topic by allowing them to convey their knowledge and ideas using their own words (Jackson & Zapata-Rivera, 2015). Students can convey their understanding in their own words in adaptable and suitable ways since conversations are iterative. CBA efficiently leverages content through conversations with

students and subsequent interactions to target specific information that may be absent from their initial responses (Jackson & Zapata-Rivera, 2015). They can guide learners on what to do next, ask questions, provide hints to elicit extra or missing information, repeat or rephrase questions, hold social interactions and provide feedback on the quality of responses through the flow of conversation. CBA can be designed using selected-response (e.g., multiple-choice and true-false items) and constructed-response (e.g., open-ended and short-answer items) formats.

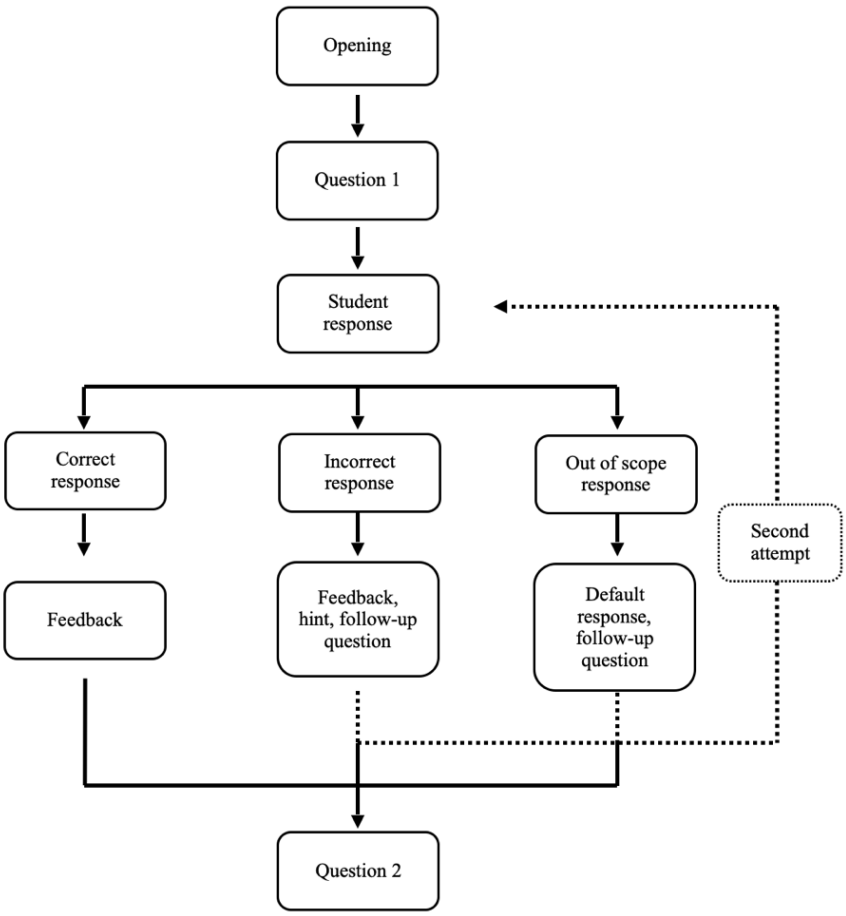
Conversation-Based Assessment with Constructed-Response Items

Constructed-response items are commonly used in educational assessment to elicit student responses that reflect underlying knowledge or skills as students must recall knowledge and compose their responses. On the other hand, students can incorporate off-topic or missing information in their answers. Thus, according to previous research, students should be given a second chance to answer questions to promote learning (Attali & Powers, 2008). CBA can ask questions, provide feedback, give hints to elicit additional or missing information, and allow a second attempt for an initial incorrect response with constructed-response items. Similar to a conventional constructed-response item, CBA starts conversations with a question (see Figure 1). Conventional digital assessments would come to an end after students respond to each question. Within a CBA, however, the student responses to that question are examined and sorted into one of several subdivisions to guide the students by feedback, hint, or follow-up question: (1) correct conversation path, where a student initial response is correct, (2) partial correct conversation path, where a student initial response is not correct but the final response is correct (3) incorrect conversation path, where both initial and final responses are incorrect, or (4) out of scope response path, where CBA is unable to determine whether their response is correct or incorrect

(see Figure 1). Out of scope response occurs when a student says something irrelevant (i.e., unrelated to the question) or gives unexpected answers that are on-topic but cannot be classified.

Figure 1

Conversation Diagram in CBA with Constructed-Response Items



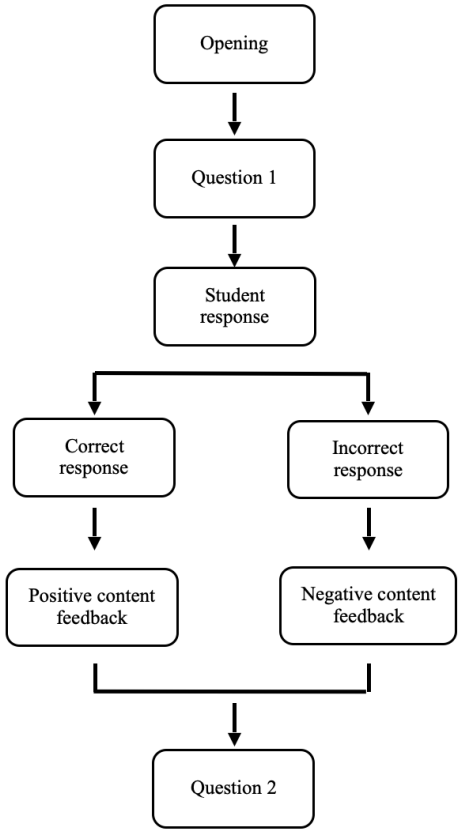
Conversation-Based Assessment with Selected-Response Items

CBA can ask questions and provide feedback in a turn-taking format with selected-response items. Thus, it combines assessment and feedback to provide real-time assessment and feedback to students, resulting in an assessment that helps students understand what they know and need to study more. In CBA with selected-response items, the student response to the first question is sorted into correct or incorrect response paths, and CBA shows the relevant feedback

and then the next question: (1) correct conversation path, where a student response is correct, (2) incorrect conversation path, where a student response is incorrect (see Figure 2).

Figure 2

Conversation Diagram in CBA with Selected-Response Items



Research Aims and Questions

The learning environment should offer affordances to students to experience cognitive disequilibrium, but also should scaffold by human tutors or computer agents. Once students are in cognitive disequilibrium, they can achieve learning with some scaffolding. At that point, cognitive equilibrium returns, and students can resume with determination, renewed curiosity, and enthusiasm when confusion is alleviated (D'Mello et al., 2007). Through their congruence with learning models, CBA has promising potential and the ability to extend computer-based

assessment and feedback and favorably impact teaching and learning (Jackson & Zapata-Rivera, 2015).

CBA has the potential to advance existing computerized formative assessments and create an interactive assessment environment, however, these systems have yet to become a standard feature of classrooms. Despite the vast research on instruction integrated conversational agents, the recent efforts to investigate and harness methods for modeling conversations for assessment purposes, and thus scientific evidence and our knowledge are limited and incomplete. To address the limited research on CBA, this study introduces a CBA with selected-response (multiple-choice and true-false) and constructed-response (short-answer) tests as a formative assessment tool for students and investigates student attitudes toward CBA to answer the following questions:

1. What is the functionality of CBA in interpreting student responses accurately?
2. What are the student attitudes toward taking an assessment with CBA?
3. What are the usability indicators and issues of CBA?

The study includes five chapters: introduction, literature review, methodology, results, and discussion. The introduction explains the motivation for the study and the theoretical framework and provides information on the aim and research questions. Chapter 2, literature review, discusses the body of the existing research surrounding the conversational agents and CBA, and their benefits and limitations in order to develop a clear understanding. Chapter 3, methodology, outlines the current research study, presents and explains the chosen methodology, the sample used, data collection, as well as data analysis. Chapter 4, results, presents the results from conversation data (research question 1) and survey data (research question 2) collected by participating students of this study and cognitive walkthrough reports (research question 3)

written by students who evaluated the usability of CBA. Chapter 5, discussion, summarizes and discusses the main results with regard to the research questions, reflects on the results, presents practical implications, strengths, and limitations of the study, and provides recommendations for future research.

Chapter 2: Literature Review

This chapter reviews, critique and discuss the existing relevant literature surrounding the conversational agents and conversation-based assessment (CBA) to investigate and understand how other researchers have approached the issue related to the current study as well as their results and conclusions. With a few exceptions, all of the studies covered in this chapter were conducted on conversational agents as the literature on CBA is limited (Jackson et al., 2018; Lopez et al., 2021; Ruan et al., 2019). Thus, this chapter endeavors to examine the literature relevant to conversational agents while focusing on a narrower field, CBA. The literature review presents the concept of conversational agents and CBA with a set of examples as well as their benefits, limitations, and designs.

Conversational Agents

The research into technology-based support in education has been motivated by the increasing demand for diverse and personalized educational needs. The intelligent tutoring system (ITS) was one of the first attempts in this regard. In the late 1970s, ITS was introduced to create computerized learning environments that could optimize each student learning (D'Mello & Graesser, 2013). ITS blends instruction and assessment for instructional purposes to adaptively respond and give immediate feedback to student responses and guide them on what to do next (Graesser et al., 2014). Research showed that ITS applications helped students learn more effectively while outperforming both computer-based training applications and novice human tutors (Corbett, 2001; Dodds & Fletcher, 2004; Wisher & Fletcher, 2004). Although ITS provided great potential for personalized education, they still considered learners as passive knowledge recipients and did not bring the notion of interaction within the learning context. To address this limitation, researchers have attempted to develop more advanced systems that could

interact with students in a more natural way by simulating human teachers (i.e., conversational agents).

It is recommended to increase interaction between the teacher and the student to improve student learning (Goel & Polepeddi, 2016); however, it is not feasible to increase interactivity with each student in a classroom where there is a large number of students. One feasible strategy to provide learning assistance to all students is to use conversational agents. Conversational agents provide a solution to this problem by simulating human teachers to increase student learning and motivation. This could free human teachers to have more time to engage in deeper discussions with their students. Conversational agents have three main components: a user who wants to achieve specific goals, actions that need to be completed to reach those goals, and a computer system with which the user can communicate to complete the actions (Maedche et al., 2019). A conversational agent can produce a dialogue (i.e., a conversation between a student and an agent) or a trialogue (i.e., a conversation between multiple students and multiple agents) with students (Davis, 2018).

Conversational agents have been studied as part of previous work on ITS and proved to be an effective learning process (e.g., AutoTutor; Graesser et al., 2004). They can be viewed as an enhanced form of ITS because conversational agents provide a learning experience with natural language dialogues while ITS does not. Conversational agents are capable of providing tailored support to each student, as well as building on each student strengths, interests, and abilities to improve engaged and independent learning (Kerly et al., 2008a). Because conversation is a channel through which nearly all students are accustomed to expressing themselves, the utilization of dialogue is a major component of conversational agents (Kerly et al., 2008a). This allows students to focus on the learning task rather than being strained by the

communication medium (Beun et al., 2003). Previous research has shown that the interactive structure of conversations creates an ideal environment for information exchange and reveals student knowledge (Graesser et al., 2008).

Although these agents are mainly designed for instructional purposes—such as virtual teaching assistants, tutors, and peers (or learning companions)—they have a wide range of application possibilities in the field of education. There are now efforts underway to investigate and harness methods for modeling conversations for assessment purposes (i.e., conversation-based assessments: CBA). CBA creates an interactive assessment environment where assessment takes place between a student and a computer agent. It can measure student learning and provide feedback through automated or adaptive conversations by a computer agent. CBA combines assessment and feedback in an optimal formula to improve student learning while assessing student knowledge and providing timely feedback. Thus, it advances computer-based assessment and feedback by simulating human teachers to monitor and enhance student learning through interactivity which is often missing in computer-based assessment and feedback.

Examples of Conversational Agents

To date, conversational agents have been utilized mainly for instructional purposes. Recently, researchers have studied the potential to expand instruction integrated conversational agents for assessment purposes (i.e., CBA). The findings supported the use of CBA to measure student knowledge and skills in a conversational environment (e.g., English language skills of second language learners; Lopez et al., 2021). The literature on conversational agents in education is examined in this part, demonstrating significant research and development effort focused on the use of conversational agents for instructional purposes in the educational context (i.e., tutoring).

There are different conversational agents that vary in the extent to which they simulate human dialogue mechanisms. All aim to comprehend natural language, formulate adaptive responses, and implement pedagogical strategies to help student learning (see Table 1). The main types of conversational agents are speech-based or text-based agents (or chatbots) that receive input from users and deliver output to them through natural language processing (NLP) (Maedche et al., 2019). Text-based forms are generally used to facilitate conversations where students type answers and questions (e.g., QuizBot; Ruan et al., 2019; see Figures 3 and 4). Speech-based forms can employ embodied conversational agents that convey emotion and gestures, as well as speech synthesis (text-to-speech) and speech recognition (speech-to-text) that enable voice input and output (e.g., ARIES; Cai et al., 2009; see Figure 5).

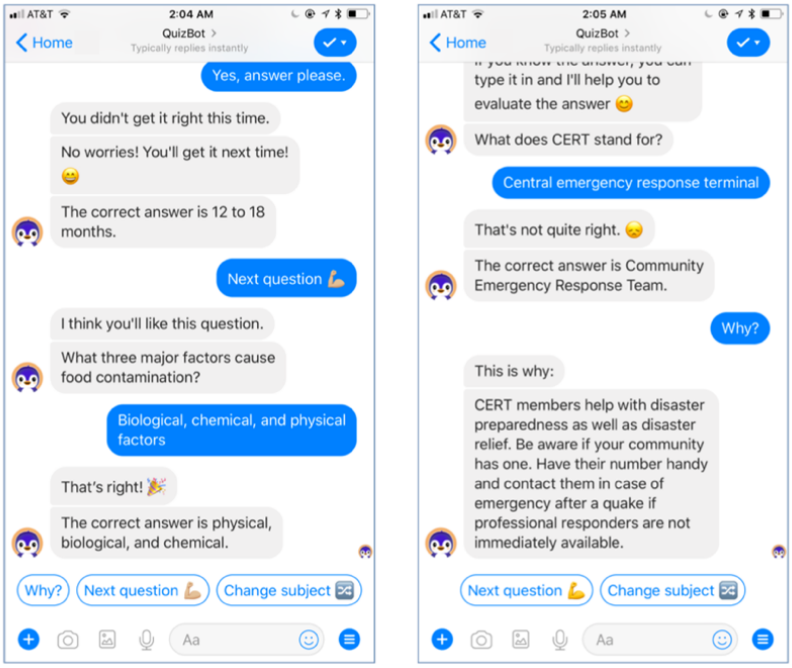
Table 1

Conversational Agents by Subject, Grade, and Purpose

	Subject focus	Grade	Purpose
AutoTutor	Computer literacy	College	Tutoring
Ms. Lindquist	Algebra	College	Tutoring
Geometry Explanation	Problem-solving in geometry	High school	Tutoring
iSTART	Reading comprehension for science	College	Training
CALMsystem	Science	Primary school	Tutoring
MetaTutor	Biology	High school	Tutoring
ARIES	Scientific inquiry	College	Tutoring
EER-Tutor	Introductory database course	Undergraduate	Tutoring
The Request Game	English	College	Tutoring
Affective AutoTutor	Computer literacy	College	Tutoring
Beetle-2	Basic electricity and electronics	College	Tutoring
DeepTutor	Science	College	Tutoring
KSC-PaL	Computer Science	Undergraduate	Tutoring
Rimac	Physics	High school	Tutoring
QuizBot	Science, safety, English vocabulary	College	Assessment
ELLA-Math	English and math	Middle school	Assessment

Figure 3

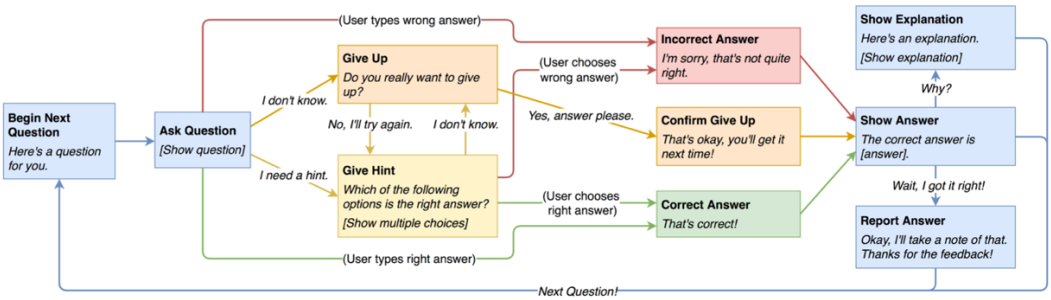
Student Types Correct (Left) and Incorrect Answer (Right) in QuizBot



Note. Reprinted from "QuizBot: A dialogue-based adaptive learning system for factual knowledge", by Ruan et al. (2019), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* p. 1. Creative Commons Attribution.

Figure 4

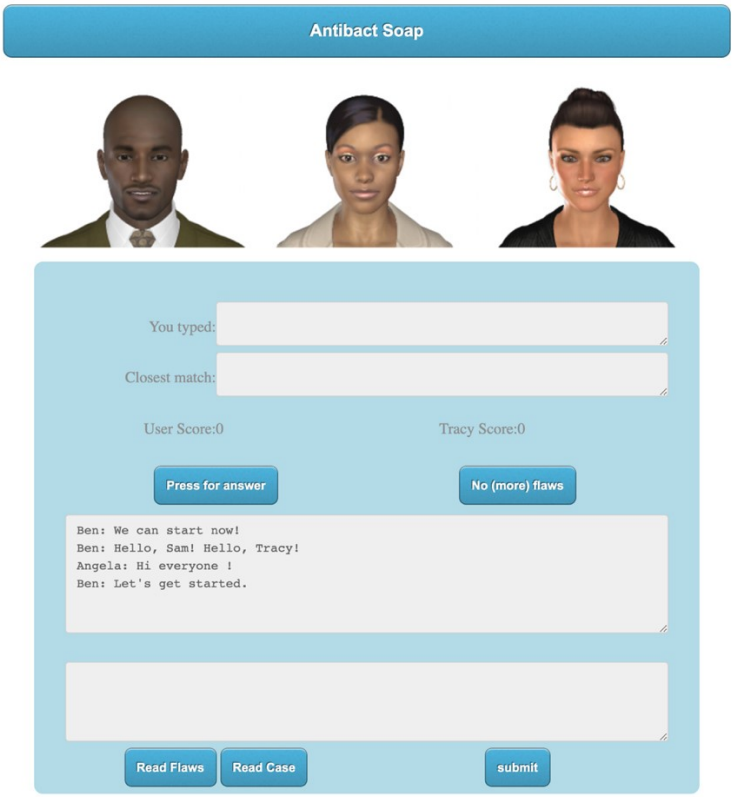
Conversation Architecture of QuizBot with Sample Responses



Note. Reprinted from "Quizbot: A dialogue-based adaptive learning system for factual knowledge", by Ruan et al. (2019), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* p. 3. Creative Commons Attribution.

Figure 5

Speech-Based Conversational Agents in an ARIES Trialogue



Note. Source: http://ace.autotutor.org/IISAutotutor/index.html

AutoTutor (Graesser et al., 1999) project was launched in 1997. It is augmented by three-dimensional interactive simulations including conversations among two animated conversational agents (a tutor and student) and students to enhance motivation and learning (Graesser et al., 2005; Jackson & Graesser, 2006). AutoTutor holds conversations with students in natural language and simulates the dialogue moves and pedagogical strategies of human tutors (Graesser et al., 2004, 2008). It was grounded in explanation-based constructivist learning theories, intelligent tutoring systems, and empirical research on dialogue patterns in tutorial discourse (Graesser et al., 2014). It presents challenging problems (or questions) and then engages in dialogue that coaches students in actively constructing an answer. There is a multi-

turn tutorial dialogue between AutoTutor and the student to answer a question (or solve a problem). That is, it assists students in constructing their answers after their initial responses.

Ms. Lindquist (Heffernan, 2003) provides assisted practice in algebra by scaffolding learning via doing instead of direct instruction. Ms. Lindquist was a "less is more" technique in which students solved fewer problems but learned more per problem when they were involved in conversations.

In **Geometry Explanation Tutor** (Aleven et al., 2001; Aleven et al., 2004), students explain their solutions to geometry problems in their own words. It classifies a student input in a hierarchical domain that contains both complete and incomplete explanations. The classification is used as a proxy for a student ability to provide thorough and accurate explanations. When the knowledge-based strategy fails, it employs a statistical text classifier to determine whether an explanation is correct or partially correct. The tutor gives scripted feedback that corresponds to the appropriate response category.

iSTART (McNamara et al., 2004) is a reading strategy trainer. It discusses and provides feedback about reading strategies through animated agents to improve reading comprehension. It tailors student-tutor interactions based on a thorough comprehension of the student contribution. Researchers showed that it facilitated both reading strategies and comprehension.

In the **CALMsystem** (Kerly et al., 2008b), students respond to questions on the topic and express their confidence in their abilities to accurately answer each question to encourage self-directed learning and the development of metacognitive skills. This creates a model of their self-assessments alongside the system inferences about their knowledge. Based on their responses, the system infers a knowledge level for them and encourages them to engage in a dialogue to reflect on their self-assessment and any discrepancies between their confidence and their

knowledge levels provided by the system. This enhanced accuracy of self-assessment substantially more than a reflection based solely on system visual inspection and student confidence (Kerly & Bull, 2008). Students who experienced the CALMsystem with the conversational agent improved the accuracy of their self-assessments and reduced the disparity between their own and the system assessments significantly compared to students who experienced the system without the conversational agent (Kerly et al., 2008b).

MetaTutor (Azevedo et al., 2009) monitors, models, and supervises students' metacognitive processes while learning a complicated scientific subject in hypermedia contexts. It was developed based on considerable research that demonstrated the importance of metacognitive activities such as goal setting and planning during the learning process to automate the function of an external agent while students learn with hypermedia. It is intended to fade into the background over time, allowing students to finally control their learning.

ARIES (Cai et al., 2009) uses two animated pedagogical agents (guide agent and student agent) to teach scientific reasoning and critical thinking abilities. The guide agent is an expert on scientific research and serves as a competent tutor. The student agent displays misunderstandings and insufficient knowledge, which the guide agent and human student can remedy.

To diagnose student solutions to database design assignments, **EER-Tutor** (Weerasinghe et al., 2009) examines a hierarchy of errors that students make. It gives adaptive feedback in the form of prepared dialogues associated with each error type.

The Request Game (Yang & Zapata-Rivera, 2010) aids second language acquisition by giving learners a stress-free environment in which to practice conversational skills. Language learners practice by sending requests to a virtual agent with the help of a dialogue engine. Based

on the acceptability of the request given by the human learner, the virtual agent offers both verbal and non-verbal (facial expressions) feedback.

Affective AutoTutor (D'Mello & Graesser, 2013) detects and responds to student emotional states in addition to the individualized instruction and human-like interactivity. Affective AutoTutor monitors facial features, body language, and conversational cues and regulates negative states such as frustration and boredom. It responds with an affective statement accompanied by a matching emotional facial and vocal expression. Thus, it detects and helps regulate negative emotional states to increase engagement and learning. It showed more dramatic improvements in student learning gain compared to the AutoTutor.

Beetle-2 (Dzikovska et al., 2014) reacts to mistakes students make while completing problem-solving activities. It compares the student explanations and other information with reference responses to analyze the discussion status across numerous turns.

DeepTutor (Rus et al., 2013, 2015) provides macro-adaptation (i.e., appropriate tasks for a student) and micro-adaptation (i.e., feedback and support at each task step) using a framework called a learning progression matrix. It is a hierarchical structure that models student performance in each course topic and across a sequence of increasingly challenging topics covered in the course. A course consists of topics, which are addressed through a series of lessons. Each lesson includes a series of tasks (e.g., problems, activities), which are accomplished through a series of solution steps and each step can be facilitated through a series of tutoring tactics (e.g., hints, prompts).

KSC-PaL (Howard et al., 2017) is a collaborative problem-solving dialogue agent in computing science education. It interacts verbally, graphically, and in a process-oriented form with a student in a one-on-one problem-solving as a peer collaborative agent. Its goal is to track

student collaborative behavior and try to steer them in the direction of more productive behavior. The peer agent appears to encourage students to examine the contributions more thoroughly, but it does not leave them to struggle for a long time when they are confused.

To deliver adaptive instruction**, Rimac** (Katz et al., 2021) dynamically constructs a student model that drives reactive and proactive decision-making to enhance the understanding of concepts associated with quantitative physics problems. Rimac asks questions throughout tutoring to make tutoring more adaptive and efficient. The student model in Rimac enables the chatbot to mimic the adaptive scaffolding provided by human tutors by selecting appropriate content to focus on (i.e., domain contingency) and addressing this content with an appropriate amount of help while a student performs a task (i.e., instructional contingency).

QuizBot (Ruan et al., 2019) helps students learn factual knowledge in science, safety, and English vocabulary. The interactions with QuizBot are a combination of typing and button options. If the user knows the answer, they can type or select, hit the "Hint" button, or tap the "I don't know" button. When a user inputs and sends an answer to QuizBot, the chatbot evaluates the response for accuracy. They can get a quick explanation by tapping the "Why" button.

ELLA-Math (Lopez et al., 2021) was created to use small-group activities to measure English learners' English proficiency and math knowledge. Students communicate with three virtual agents: a teacher and two student agents, but the majority of the interactions are between the student and the two student agents.

Benefits of Conversational Agents

Learning and motivation have been facilitated via conversational agents incorporating virtual agents and a human student. AI-based digital assistants open up new possibilities for increasing levels of motivation, feedback, and personalization in the learning process, and hence

learning outcomes (Maedche et al., 2019). Personalization is especially vital for efficient learning in order to adapt to students from various backgrounds (McAndrew & Scanlon, 2013). Early research reveals that AI-based assistants can play a role similar to a human tutor in helping learners improve their task performance and skill levels over time, especially when it comes to abilities like problem-solving (Winkler et al., 2019).

Student Learning

Studies reported that AutoTutor produced statistically significant learning gains depending on the comparison condition and the version of AutoTutor (Graesser et al., 2005). AutoTutor improved student average learning gains nearly one letter grade compared to the reading textbook for an equivalent amount of time (Graesser et al., 2005, 2008; Nye et al., 2014). It was also effective on student average learning gains for deep levels of comprehension in comparison with (1) reading nothing, (2) starting at pretest, or (3) reading the textbook for an amount of time equivalent to that involved in interacting AutoTutor (Graesser et al., 2014). Another study compared AutoTutor with a condition where the course textbook was assigned to read for an equivalent amount of time to AutoTutor and with a condition where no reading material was assigned to read (Graesser, Jackson et al., 2003). AutoTutor produced significantly better learning than the two comparison conditions. In addition, a comparison of AutoTutor and novice human tutors showed that the student average learning gains were virtually equivalent on the same topic (VanLehn et al., 2007).

Researchers contrasted a conversational agent to a flashcard app, both of which used the same algorithm, same question selection model, question pool, hints, and explanations, across science, safety, and English vocabulary (Ruan et al., 2019). They discovered that when students used the chatbot, they gave more correct responses to factual knowledge questions for all three

subjects. The agent was substantially more effective in assisting people in the recall and recognition of factual knowledge. They also suggested that the effectiveness of conversational agents may generalize to different domains to measure factual knowledge—such as biology and history—as they found success in all three subjects. Another study showed that students who interact with a peer collaborative agent improve their knowledge even though there was no difference in learning gains when comparing two agent versions that do and do not track and attempt to change the student collaborative behavior (Howard et al., 2017). In comparison to constructed-response items, researchers investigated the potential benefits of CBA and found that when compared to constructed-response items, CBA items allowed 41% of students to submit a more thorough response and enhance their scores (Jackson et al., 2018). Researchers also compared CBA with multiple-choice designs and found that students who explained by conversation learned better how to provide general explanations for problem-solving steps than those who explained by choice selection (Aleven et al., 2004).

Student Motivation

Practicing with conversational agents has been found more time-consuming compared to a flashcard app (Ruan et al., 2019). Students, however, stated that the virtual agent was more beneficial for learning and chose to spend more time with the agent when given the option although it was more time demanding. Ruan et al. (2019) suggested that the conversational agents are more engaging to use and less efficient in terms of time spent, however, students can still prefer them as they enhance learning and motivation. Another study revealed that conversation was effective in keeping students motivated and had a strong positive impact on student motivation (Heffernan, 2003). Moreover, the researcher found a strong positive impact on learning and reported that students who used a conversational agent solved fewer problems

but learned as well as or better than students who were simply given the solution. This finding has been characterized as “less is more”. In another study, students found conversational environments as an engaging and easy way to practice and learn English as a second language (Hong et al., 2014; Yang & Zapata-Rivera, 2010). Students who interacted with the digital agent were shown to be more actively engaged in learning activities and outperformed those who did not (Hong et al., 2014). Moreover, students expressed an interest in using digital agents in their other subjects. In a particular study, researchers also investigated the emotional states of students when interacting with conversational agents (D’Mello & Graesser, 2013). Among different emotion states, engagement (or flow) has been found the most frequent state followed by boredom and confusion. Another study also revealed a significant relationship between learning and the affective state of flow (Craig et al., 2004).

Feedback

To obtain a better understanding of one-on-one tutoring, Graesser and his colleagues (1994, 1995, 1999) videotaped and analyzed tutoring sessions. They found that tutors give positive feedback rather than negative feedback to student misconceptions or incorrect answers. Instructors can fail to give negative feedback when students perform poorly as they tend to follow politeness in their conversations (Graesser et al., 1995; Ogan, Finkelstein, Mayfield et al., 2012; Ogan, Finkelstein, Walker et al., 2012; Wang et al., 2012). Conversational agents have the potential to provide optimal feedback to students by satisfying the trade-off between feedback accuracy and politeness.

Different versions of AutoTutor were constructed by Jackson and Graesser (2007) to control the feedback that college students received throughout their interactions with the agent. Content feedback (e.g., providing key information in student response), progress feedback (e.g.,

evaluating their performance), both or none were given to students. Although students benefited in all of the feedback settings, content feedback had a stronger influence on learning than progress feedback.

Even though instructors often assume that students can understand the feedback given (i.e., feedback literacy), students may not be able to understand the feedback. Researchers investigated the role of feedback in conversational agents and found that when students interact with an agent, they are under the impression that the agent cares what the student communicates (Graesser et al., 1999). Previous research showed that most students appreciated how well the agents asked follow-up questions and provided guidance and feedback to help them comprehend the questions (Lopez et al., 2021). It has been suggested that feedback helps enhance the testing effect in CBA regardless of whether the attempted answers are correct or not (Ruan et al., 2019). Thus, feedback can be more motivating and encouraging to learn in a CBA environment. To increase feedback literacy, the agents can enhance active learning by asking questions about student understanding of the given feedback.

Limitations of Conversational Agents

A conversational agent is still a challenging work in progress and implementing a conversational agent in practice is more difficult than expected (Jackson & Zapata-Rivera, 2015; Yu et al., 2017). We must be cognizant of the current limitations. Simple activities, such as scheduling an appointment based on an e-mail request, might be completed by an agent with minimal human participation. However, more difficult activities or decisions may require more human participation. For example, CBA is not yet well-developed to comprehend content at the level we consider necessary and to grade the content quality of student answers (e.g., Maedche et al., 2019).

Irrelevant Responses

Previous research reported some challenges in conversational agents. One of the main problems was that conversational mechanisms in automated environments could not handle most of the student questions and provide relevant and correct answers as student questions are mostly unpredictable to write every possible pattern (Graesser, 2016). Conversational agents can correctly answer only a modest proportion of student questions (e.g., AutoTutor) and students may become frustrated when breakdowns (e.g., unresponsiveness to what the student says) occur (Bailey et al., 2021; D'Mello & Graesser, 2013). Another study found that learners were dissatisfied with the responses provided by a foreign language practice agent because of insufficient pattern-matching mechanisms (Jia, 2003). According to previous research, although conversational agents function well with selected responses (e.g., multiple-choice) or short responses (e.g., typing yes or no), they do not function well with open-ended responses (Huang et al., 2019; Valério et al., 2018). Simple communication alternatives, such as simple short responses or button response options, are suggested by researchers to minimize irrelevant responses (e.g., Valério et al., 2018). Lopez et al. (2021) reported that CBA was more accurate in interpreting student responses to questions that require writing numbers rather than writing words as well as shorter responses (e.g., numbers and one or two words) rather than longer responses. In addition, CBA is more accurate if a response requires more flexible answers (e.g., multiple synonyms for *summative assessment*) rather than more specific words (e.g., *criterion-referenced* or *absolute grading*).

Inaccurate Feedback

Conversational agents are generally designed to give positive feedback for a more complete answer (e.g., AutoTutor). That is if the student answer is partially correct the agent still

provides negative or neutral feedback for their partial answer, however, this can frustrate students (Graesser, 2016). Furthermore, conversational agents can provide negative feedback if a correct answer is matched with negative feedback or the opposite scenario. For example, even if the initial response is correct, the system may occasionally convey responses that contain misspelled words to other discussion pathways. If this happens, students may receive inappropriate or irrelevant follow-up questions, hints, or feedback (Lopez et al., 2021). In their research, Lopez et al. (2021) found that almost half of the students reported that the system did not always understand their answers. The inaccurate feedback can also confuse, demotivate and frustrate students (Graesser, 2016; Lopez et al. 2021). To deal with inaccurate feedback in conversational agents, the researchers suggest using neutral short feedback rather than negative or positive long feedback with the hope that if mismatches occur in the system students will not be demotivated or frustrated (Graesser, 2016).

Excessive Interaction

Researchers also investigated the ideal amount of interaction with conversational agents. Student interviews and surveys showed that many students regarded dialogues in conversational agents to be too protracted (Katz et al., 2021). Researchers found that students judged the conversational agents to be unsatisfactory and unhelpful in providing feedback since they may spend more time on concepts that the student already understands and less time on concepts that they struggle with. Studies compared two versions of AutoTutor: (1) the control version engaged students in dialogues about six conceptual questions, while (2) the experimental version engaged students in dialogues about three conceptual questions and then presented three additional questions but did not engage students in dialogues and instead provided a predetermined short response and explanation (Kopp et al., 2012). The results showed that students in the

experimental condition learned as much as students in the control condition in less time, implying that a large amount of interaction is not always required.

Another study (Jordan et al., 2016) compared two versions of Rimac: (1) the control version that always decomposes a step to its simplest sub-steps regardless of the student knowledge level, and (2) the experimental version that adaptively decides to decompose a step based on student knowledge. The results showed that students who used the experimental version learned similarly to those students who used the control version yet spent less time. These results suggest that students may become frustrated if they believe the agent is forcing them to engage in lengthy talks about a subject that they already know rather than tackling content that they require assistance with (e.g., Kopp et al., 2012; Jordan et al. 2018). To overcome this problem, more knowledgeable students who provide correct answers can be permitted to proceed to a more difficult problem, while less knowledgeable students can receive the assistance they require. Also, there are conversational agents where emotional support is embedded (e.g., AutoTutor). However, emotional support may not motivate students because most students aim to learn the content rather than obtain emotional support (D'Mello & Graesser, 2013).

Medium of Communication

Suppose the mediums of communication used by the chatbot and student are different (e.g., in the AutoTutor, student type their responses and the chatbot speaks). In that case, this can confuse, demotivate and frustrate students (D'Mello & Graesser, 2013). Even though learning gain was positively associated with confusion and flow it was found to be adversely related to boredom (Craig et al., 2004).

In summary, previous studies showed that conversational agents could correctly answer a modest proportion of student questions, deliver inaccurate feedback that can confuse, demotivate

and frustrate students, and provide lengthy talks. These challenges, limitations, or problems still exist in CBA despite advances in computers, technology, NLP, and AI. For example, even though AutoTutor has existed for more than two decades, and the researchers continue to develop and update the system, these problems still occur as a result of the vagueness of the language.

Designing a Conversational Agent

Four important parts need to be considered to design a successful conversational agent: type of the conversational agent, subject, knowledge of the learner, and sophistication of the dialogue strategies. In terms of the first part, the type of the conversational agent, Graesser and his colleagues investigated the features of AutoTutor that might account for improvements in learning with conversational agents (Graesser, Moreno et al., 2003; Graesser et al., 2004, 2008; VanLehn et al., 2007). Their experiments showed that most of the improvement was because of the dialogue content of what the agent says not the speech or animated facial display. Thus, their findings suggest what has been expressed in a conversation matters to promote student learning. This highlights that the medium does not convey the message; the message itself is the message.

Regarding the subject, the literature shows mixed results. For example, AutoTutor worked better when the content was for qualitative domains (Graesser et al., 2005). However, another study reported that ELLA-Math was better at reading responses to questions that needed students to write a number than questions that required students to write words (Lopez et al., 2021). However, the existing conversational agents have been mainly designed for verbal or qualitative content rather than numerical or quantitative content (see Table 1). Regarding knowledge of the learner, CBA could work better when they are designed for students with low to medium levels of knowledge rather than students with a high level of knowledge (Graesser et

al., 2005). When students with a high level of knowledge interact with a CBA, it has been found that both dialogue participants (i.e., the agent and the student) expect a higher level of precision and this can lead to a higher risk of failing to meet expectations of both participants (Graesser et al., 2005). For example, AutoTutor worked better when the shared knowledge between the agent and learner is low or moderate (Graesser et al., 2005).

In terms of the sophistication of the dialogue strategies, Graesser and his colleagues (Graesser & Person, 1994; Graesser et al., 1995) investigated tutoring strategies by analyzing novel human tutors and found that they rarely used sophisticated tutoring strategies instead they tend to guide students based on expectation and misconception tailored dialogue (EMT dialogue), which is known to be common in human tutoring (Graesser et al., 2005), but still their strategies were effective. According to EMT, human tutors typically have a list of anticipated correct answers (called expectations) and a list of anticipated incorrect answers (called misconceptions) associated with each question or problem (Graesser et al., 2005; Lopez et al., 2021). They ask questions that are within student ability to answer correctly, compare student input with their anticipated expectations and misconceptions, and then provide support when students need to avert an incorrect answer.

Current Study

CBA advances conventional digital assessments by simulating human teachers to increase student learning and motivation through interactivity and assistance that are often missing in digital assessments. CBA can provide personalized help to each student while also assessing their learning. They can build on each student's strengths, interests, and abilities to enhance learning and motivation. Through the natural flow of conversation, they can hold social interactions with students, ask questions, provide hints, direct students on what to do next, and

provide feedback on the quality of responses. Despite the aforementioned mounting evidence that conversational agents help student learning and motivation, these systems have yet to become a standard feature of higher education classrooms. In addition, despite the recent efforts to investigate and harness methods for modeling conversations for assessment purposes, most conversational agents are instruction integrated and thus they are designed for tutoring purposes (see Table 1). Thus, scientific evidence and our knowledge of CBA are limited and incomplete. To address this gap and contribute to the literature on the utility of CBA in monitoring student learning and improving student attitudes toward taking an assessment in an interactive environment, this study aims to design and implement a new CBA for two sections of a large-size undergraduate-level course, EDPY 303 Educational Assessment, at the University of Alberta (n_1 = 290 and n_2 = 119) to answer the following questions:

1. What is the functionality of CBA in interpreting student responses accurately?
2. What are the student attitudes toward taking an assessment with CBA?
3. What are the usability indicators and issues of CBA?

Chapter 3: Methodology

This study aimed to design a CBA that can measure student knowledge and provide support and feedback to scaffold their learning. CBA was designed for two sections of an undergraduate-level course, EDPY 303 Educational Assessment, a mandatory course for all undergraduate students enrolled in the Elementary and Secondary Education programs in the Faculty of Education at the University of Alberta. CBA was offered to students as an additional and optional formative assessment tool by the course instructors in the 2021-2022 academic year. Table 2 presents the background information about the participating students in the survey (n = 61). The next page explains the research questions aimed to answer in this study by the design and approaches summarized.

Table 2

Background Information about Students

Demographic variable	Number of students
Age	
25 years or below	45
26-30 years	10
31 years or above	4
Gender	
Female	51
Male	9
Non-binary	1
Year	
Year 3	39
Year 4	8
Year 5+	14
Chatbot experience	
Yes	15
No	44
Not sure	1
Technology skill	
Intermediate	54
Expert	7
Content knowledge	
Not confident	2
Somewhat confident	38
Confident	21

1. **What is the functionality of CBA in interpreting student responses accurately?**

CBA consists of two constructed-response (see Appendix A) and three selected-response tests (see Appendix B) following the previous research that designed conversational agents with both formats (e.g., Lopez et al., 2021; Ruan et al., 2019) and also the preference of the course instructors. Table 3 shows further details about each test including the availability period as well as the number of items in each test. Selected-response tests combine assessment and feedback to measure student knowledge and provide timely feedback. The back-and-forth dialogue is intended to be a turn-taking conversation where the agent asks a question, the student responds, and the agent provides feedback and asks the next question. Constructed-response tests combine assessment, scaffolding, and feedback to measure student knowledge, give a second attempt for their initial incorrect or out-of-scope responses and provide feedback. CBA with the selected-response tests was available for both sections of the course (n_1 = 290 and n_2 = 119), while CBA with the constructed-response tests was available for only the second section (n_2 = 119) following course instructors' availability and preference to use CBA in their sections. The unique total number of students who took selected-response and constructed-response tests are 98 and 21, respectively. Table 4 shows the number of participating students in each test. Conversation data from each test was used to calculate the intent classification and confidence score to investigate the functionality of CBA with constructed-response and selected-response tests in interpreting student responses accurately.

2. **What are the student attitudes toward taking an assessment with CBA?**

CBA sent students a self-assessment question and asked them to evaluate their own performance once they completed the assessment (see Appendix C). The purpose was to provide general support to students based on their own evaluation of their performances. Following the

self-assessment question, CBA sent a survey link and invited them to complete an experience survey (see Appendix D). As indicated above the unique total number of students is 98 for selected-response and 21 for constructed-response format. However, the unique total number of students who completed the survey is 61. Thus, even though more students experienced CBA relatively small group of the participating students filled out the survey, Table 4 shows the number of students in each survey, and Table 2 shows the background information about the participating students in the survey. The survey consisted of background questions related to demographic information (e.g., age, gender), technology use, and content knowledge. Students were asked a series of questions to better characterize their engagement and overall experience with CBA. For example, they were asked to score their level of agreement with statements concerning general engagement with CBA. Ethical approval was obtained from the Research Ethics Office for the use of survey data and secondary use of conversation data.

3. **What are the usability issues of CBA?**

CBA was shared with students (n_3 = 106) enrolled in CMPUT 302 Introduction to Human Computer Interaction–another undergraduate-level course at the University of Alberta focusing on a user-centered approach to software design. This course requires students to conduct cognitive walkthroughs for different software designs. The course instructor contacted to offer CBA to their students for their service-learning project. The availability period of CBA to students was decided by the course instructor. Students were grouped into 21 teams and performed the cognitive walkthrough method to evaluate the usability of CBA for potential usage scenarios (i.e., actions). They were not trained on how to use CBA and thus their cognitive walkthrough was more efficient to identify possible usability problems. Each team prepared a report answering the following questions:

- Will the users try to achieve the right effect (i.e., outcome)? This question examines if the users follow the correct path for a specific outcome.
- Will the users notice that the correct action is available? This question examines if there are confusing options that could prevent users from following the correct path.
- Will the user associate the correct action with the effect trying to be achieved? This question examines if the users will be able to understand the options and make the correct decision.
- If the correct action is performed, will the user see that progress is being made toward the solution of the task? This question examines if the users are affirmed that they are on the correct path for a task whenever they take the correct action.
- How can it be improved? This question examines actionable suggestions to fix the issues and improve the tool for all potential users.

Table 3

A Summary of the CBA Designs

	Availability	Availability period	Number of items
Selected-response test 1	Sections 1 and 2	January 17-20	8
Selected-response test 2	Sections 1 and 2	February 3-10	7
Selected-response test 3	Sections 1 and 2	February 25-March 8	8
Constructed-response test 1	Section 2	January 21-24	3
Constructed-response test 2	Section 2	February 17-22	4

Table 4

The Number of Students in CBA and Survey by Each Test

	CBA	Survey
Selected-response test 1	67	40
Selected-response test 2	77	22
Selected-response test 3	58	14
Constructed-response test 1	19	3
Constructed-response test 2	7	0

Design of CBA

Following the aforementioned parts in the design of a conversational agent (i.e., type of the conversational agent, subject, knowledge of the learner, and sophistication of the dialogue strategies), CBA was designed to be a chatbot (i.e., text-based assistant) rather than a speech-based assistant to assess student knowledge and convey the message. Thus, both students and the agent communicated using the same medium (i.e., text). The aim was to avoid the problem of being confused, demotivated, and frustrated due to the different mediums of communication. Despite the mixed results in the literature regarding the conversational agent on quantitative versus qualitative subjects (e.g., Graesser et al., 2005; Lopez et al., 2021), qualitative content was adapted into CBA considering the high number of successful conversational agents on qualitative subjects. Also, the shared knowledge between CBA and students was satisfied with developing CBA through conversation with course instructors with the most relevant information. The aim was to deal with the limitation of unexpected student input that can lead to unresponsiveness to what students say. Finally, this study followed the EMT dialogue, and thus, the dialogue mechanism in CBA is computationally manageable and similar to what human tutors do: CBA asks a question and assesses the student knowledge based on a list of expectations, provides hints, corrects misconceptions, and gives feedback. The expectations and misconceptions associated with each question were stored in the CBA script (i.e., the major content repository of questions and dialogue moves).

CBA Script

CBA script included questions, correct answers (expectations), incorrect answers (misconceptions), hints, prompts, feedback, and other inner loop information. For each question, there are correct and incorrect answers that students are expected to provide to a question. These

expectations and misconceptions were coded into the CBA script along with questions. CBA (1) starts with a daily conversation (e.g., students greet the agent, and the agent identifies the greeting intent and responds to the greeting) with the intention of social interaction, (2) asks questions, (3) gives hints to revise initial incorrect answers for constructed-response items, (4) provides feedback, (5) summarizes answers and (6) supports student self-assessment.

Daily conversation. Regarding the daily conversation at the beginning of CBA, it is short to meet the trade-off between engagement and efficiency because the social conversational side of CBA can lead to inefficiencies in assessment. For example, studies found that about 12 percent of the total time was not spent on learning (e.g., 4 percent of the total time was a manual delay because of the deliberate delays to make it feel like a real person, 2 percent of the total time was just chatting with the chatbot for fun; Ruan et al., 2019). Removal of these casual aspects can negatively affect student interest or motivation and thus, the trade-off between engagement and efficiency was taken into account while designing CBA.

Questions. Questions were written through conversations with the course instructors who had the most relevant information rather than guessing what topics students would struggle with and require further assistance. In addition, the goal was to address the problem of unexpected student inputs, which can lead to the unresponsiveness of the agent to what students say. It was also aimed to avoid underexposing students who do not understand the material and overexposing students who firmly understand the material. Furthermore, to deal with the limitation of positive feedback to a more complete answer (i.e., inaccurate feedback to a partially correct response) in constructed-response tests, short-answer questions were written to allow for better input recognition (Katz et al., 2021). Like most conversational agents (e.g., AutoTutor), the order of the questions was fixed.

Feedback and Support. CBA follows a particular order to select a dialogue turn to provide feedback or support. Feedback and support are built into the conversations and triggered by how CBA matches the student response. CBA provides feedback in selected-response tests and feedback and support (e.g., hint and a follow-up question to shift the conversation from the agent to the student) to allow the student to correct their initial incorrect answers in constructed-response tests. Following previous studies in the literature (e.g., Jackson & Graesser, 2007), if student response is matched with the set of expectations, CBA presents some positive short content feedback at the beginning of its next turn in order to satisfy the trade-off between feedback accuracy and politeness. In contrast, if there is a match between the student response and the set of misconceptions, CBA presents some sort of negative short content feedback at the beginning of its next turn or provides a hint (for constructed-response tests). It prompts the student in this scenario by providing them a second chance to respond. If the hint fails (i.e., the student cannot correct their inaccurate answer), CBA delivers feedback and sends the next question. Therefore, as the student expresses information over the turns, CBA compares the information with anticipated correct and incorrect answers and formulates the dialogue moves based on the student input.

Summary and Self-Assessment. CBA also gives a summary along with the feedback to recap the answer to a question before presenting the next question. There are two reasons to provide a summary: (1) students can focus on the next question and (2) students may provide a correct answer by flawed thinking or guessing, and thus it could be beneficial to explicitly cover and summarize the question before presenting the next one. However, this summary is short to avoid forcing students to engage in lengthy talks. Once students finished the assessment, they

were asked to assess their performance with the question, "*Thank you for reviewing some topics with me. How would you rate your own performance on these questions?*"

Conversation Paths

Once students select or type their answers to the questions, their inputs are compared with anticipated correct or incorrect answers to match a path in CBA and provide a solution. This is also similar to what human instructors do: they give the answers with respect to matching correct or incorrect answers to guide students. CBA controls the direction of the interaction to increase the achievement of pattern completion because it is impossible to write every possible pattern in conversational mechanisms to provide relevant and accurate answers to student questions (e.g., Graesser, 2016). To handle this, CBA guides students in their responses. For example, the agent offers an assessment to practice, but the student influences the direction of the conversation based on their input; that is, the student input influences the outputs of CBA. The aim was to overcome the problem of unexpected inputs, resulting in a CBA that functions better.

CBA includes several conversation paths based on the student input within a conversation. It follows a tree structure based on the student input (see Figure 1 for CBA with constructed-response items and Figure 2 for CBA with selected-response items). When a question is presented, the question is answered through an interaction between the agent and the student by a 4-step frame: (1) the agent presents a question, (2) the student gives an answer, (3) the agent gives feedback or hints based on the type of question (i.e., constructed-response or selected-response item) and the student answer (i.e., correct or incorrect answer), and (4) the agent summarizes the answer and shows the following question. Thus, in the 4-step frame, CBA selects an action to achieve pattern completion.

In constructed-response items, students are directed down one of four conversation paths: (1) correct conversation path: question → correct answer → feedback → next question, (2) partial correct conversation path: question → incorrect answer → hint → correct answer → feedback → next question, (3) incorrect conversation path: question → incorrect answer → hint → incorrect answer → feedback → next question, (4) out of scope conversation path: question → out of scope response → default response (see Figure 1). In selected-response items, students are directed down one of two conversation paths: (1) correct conversation path: question → correct answer → feedback → next question, (2) incorrect conversation path: question → incorrect answer → feedback → next question (see Figure 2).

Rasa Framework

Natural Language Understanding (NLU) was used in this study because it enables the agent to interpret the natural language input, while when conversational agents are trained on a corpus, they can be restricted to the domain of the corpus (Kerly et al., 2008a). Previous research examined the performance of the most commonly used NLU tools, namely IBM Watson, Google Dialogflow, Rasa, and Microsoft LUIS (Abdellatif et al., 2021). Rasa had the highest confidence scores for accurately classified intents. That is, classification is highly likely to be correct when Rasa produces a high confidence score for it. Considering the high confidence score Rasa produces, CBA was designed using Rasa to match student inputs to the list of expectations or misconceptions. Once the student provides an answer, CBA receives and compares the student input with the possible expectations or misconceptions. It calculates probabilities for each intent defined and matches it with the highest probability. A student can provide answers by misspelling or insufficient grammar, but as long as the words that they use are associated with

the listed words in the Rasa space, it makes the correct pattern match. Rasa includes two separate modules: natural language understanding (Rasa NLU) and dialogue management (Rasa Core).

Rasa NLU

The NLU extracts structured information (i.e., the intent) from unstructured student responses using machine learning and NLP approaches (Abdellatif et al., 2021). When CBA receives student response, it analyses it and responds using the NLU (Abdellatif et al., 2021). Similar to other NLUs, Rasa NLU classifies the intents from a given student response if they exist in the CBA script. Rasa NLU is modularized with pipelines that define how student responses are processed (Rafla & Kennington, 2019). To configure the Rasa NLU in the designed CBA, the following pipelines were used:

- Whitespace Tokenizer breaks text into terms.
- REGEX Featurizer creates a vector representation.
- Count Vectors Featurizer creates a bag-of-words representation.
- Dual Intent Entity Transformer (DIET) classifies intents.
- Response Selector builds a response retrieval model.
- Fallback Classifier classifies a user input as a fallback when the confidence score falls below a specified threshold.

Rasa Core

Rasa Core handles dialogue management, which entails choosing what actions CBA should take in response to student responses (Shahriar Khan et al., 2021). In CBA, Rasa Core uses rule-based and machine learning-based policies in the following order:

- Rule Policy handles conversation parts that follow a fixed behavior and makes predictions using rules that have been set. It enables responses for out-of-scope

communications for which CBA has not been trained, allowing it to fall back to a default response when confidence values fall below a set threshold. The default response in CBA is: "*I'm not quite sure what you mean by that response. It's okay to take your best guess to answer the question. Try again!"*

- Memorization Policy determines whether the present dialogue corresponds to the stories defined.
- Transformer Embedding Dialogue (TED) leverages transformers to decide which action to take next.

Data Structure in Rasa

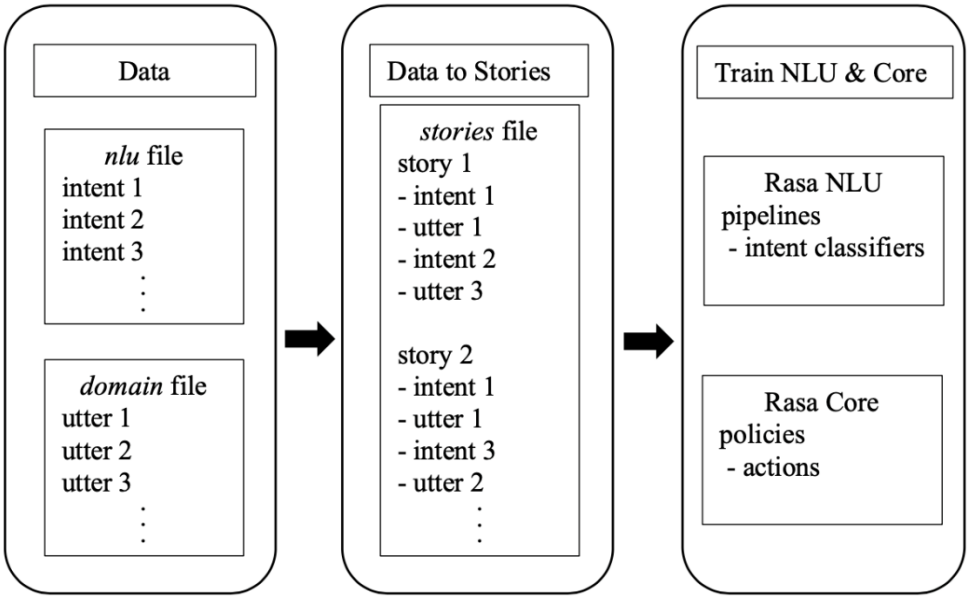
Figure 6 demonstrates the process beginning with data simulation, following with story generation, and finalizing the process with the Rasa NLU and Core training. The Rasa framework necessitates the division of the CBA script into three files (Shahriar Khan et al., 2021):

- *nlu* file is required for NLU training and contains all student responses to each question organized into intents, with each intent including different instances. A set of intents has been defined in CBA; however, as discussed in the literature, misclassification of intents might negatively influence the student experience (e.g., irrelevant response or inaccurate feedback). NLU tends to misclassify intents that more frequently share words with other intents with less exclusive words (Abdellatif et al., 2021). For each intent, the NLU was trained on a collection of student responses representing different ways a student could communicate the same response. For example, the input "formative assessment" can also be shared with "assessment for learning". These inputs were used to train the NLU on identifying the *Formative Assessment* intent.

- *domain* file contains CBA outputs (e.g., questions, feedback, and hints) to each corresponding student response in the *nlu* file.
- *stories* file contains dialogue sections, with each section containing i) a series of sequential intents that are extracted from the *nlu* file, and ii) actions that can be given when a student response is categorized under a certain intent. A story represents the conversation between a student and CBA in each chat flow.

Figure 6

Process from Data Simulation to Training



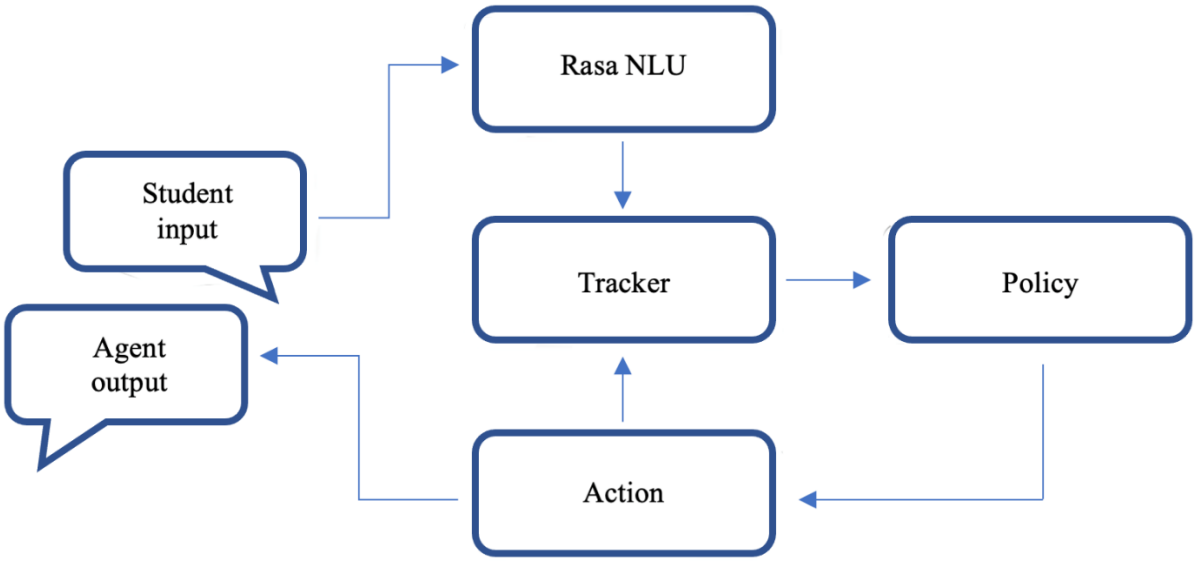
Rasa Flow

Rasa NLU and Core are fully decoupled, allowing learned dialogue models to be reused across languages and Rasa NLU and Core to be used independently of one another (Bocklisch et al., 2017). Rasa processes student responses in a series of phases, as shown in Figure 7 (Bocklisch et al., 2017). Rasa NLU performs only the first step while Rasa Core performs the rest. The NLU module takes student response (*Student input*) and converts it into a structured output that includes the original text and intents (*Rasa NLU*). The Core updates the current state

of the output from the previous state and maintains the state of the conversation (*Tracker*). Policies defined in the Core (*Policy*) use the output from the tracker to select an appropriate response from the *domain* file and execute an action (*Action* and *Agent output*). When an action is completed, it is given a tracker instance (*Action* and *Tracker*), which allows it to use any relevant information gathered throughout the dialogue history.

Figure 7

Phases from Input to Output in Rasa



Note. Adapted from "Rasa: Open source language understanding and dialogue management", by Bocklisch et al. (2017), *arXiv preprint*, p. 3.

Deployment and Pilot Study

The Rasa tool has built-in connectors that allow conversational agents to be integrated with communication platforms. After CBA was written in Rasa, it was connected to Google Chat to be delivered to students over the Internet. The trained NLU and Core modules were deployed to a hosted web server. Google Cloud Platform was used to build a connection between the hosted web server and Google Chat to chat with the trained CBA. Students had access to CBA

through Google Chat. Conversations were stored in a password-protected personal computer using an SQL database.

CBA was tested before it was considered mature enough for student use. CBA was piloted because the information on student knowledge might be inaccurate if student responses could not be appropriately interpreted and students were sent to a wrong conversation path by CBA. Therefore, before sharing CBA with students, it was tested by the course instructors and teaching assistants and tweaked accordingly.

Performance Evaluation

Intent Classification

Intent classification is the performance of NLU on correct identification of the intent of the student response. The literature suggests that there is an increase in the accuracy of the intent classification when intents contain exclusive words (Abdellatif et al., 2021). In addition, according to studies, NLUs perform better when categorizing inputs with more examples involving various ways of communicating that intent (e.g., synonyms) (Abdellatif et al., 2021). To increase intent classification, intents were created with exclusive words in the CBA script, that is, words that do not appear in other intents, and with different ways a student could communicate the same response.

Considering the binary classification of student response to each item can be either positive (i.e., classification of student response as correct) or negative (i.e., classification of student response as incorrect), true positives (TP; the number of correctly classified correct responses), false positives (FP; the number of incorrectly classified correct responses), true negatives (TN; the number of correctly classified incorrect responses), and false negatives (FN; the number of incorrectly classified incorrect responses) were calculated (see Table 5 for

indices). Using these indices, similar to previous work (e.g., Abdellatif et al., 2021), the standard classification accuracy measures—precision, recall, and F1-measure—were calculated for intent classification to evaluate the performance of CBA.

Precision is used because it provides the proportion at which positive predictions (i.e., correct responses) are correct: $\frac{TP}{TP+FP}$. Thus, precision is helpful if the number of incorrectly classified correct responses is high (i.e., FP). Recall is used as it indicates the proportion of positives (i.e., correct responses) that are correctly identified (i.e., the proportion of correct responses that are correctly identified as correct responses): $\frac{TP}{TP+FN}$. Recall is helpful if the number of incorrectly classified incorrect responses is high (i.e., FN). F1-measure is used rather than the accuracy measure because the incorrect classifications are more important than correct classifications to understand the performance of CBA in interpreting student responses. F1-measure can provide a better measure of the incorrectly classified responses as it is the harmonic mean of precision and recall: $2 * \frac{Precision*Recall}{Precision+Recall}$. However, accuracy is the measure of all correctly identified responses and it can give a better measure when class distribution is similar ($\frac{TP+TN}{TP+FP+TN+FN}$).

Table 5

Indices for Performance Evaluation

		Predicted response	
		Correct response	Incorrect response
Response	Correct response	True positives (TP)	False negatives (FN)
	Incorrect response	False positives (FP)	True negatives (TN)

Confidence Score

Another approach for evaluating CBA performance is the confidence score. The confidence score is yielded by the NLU when correctly classifying and misclassifying student responses and is scored on a scale of 0 to 1 (not confident to completely confident) (Abdellatif et al., 2021). NLU should provide high confidence scores for intents that are correctly classified while providing low confidence scores for intents that are incorrectly classified. For the confidence score, the median confidence score for the correctly classified intents of each task was used. When the confidence score falls below a specified threshold, the default response to student responses with intents out of scope directs them with the message, "*I'm not quite sure what you mean by that response. It's okay to take your best guess to answer the question. Try again!*"

Standard classification accuracy measures and median confidence scores were calculated to understand the functionality of CBA in interpreting student responses, but with slightly different purposes for each CBA format (i.e., constructed-response and selected-response). In terms of constructed-response format, these measures were calculated to evaluate the performance of CBA in understanding and processing students' written responses. On the other hand, for the selected-response format, the goal was to evaluate how accurately the CBA design was implemented. Thus, even though there were no written responses by students in CBA with selected-response format, the aim was to check the accuracy between system design and system implementation.

Analysis of Cognitive Walkthrough: From Codes to Themes

CBA with one selected-response test was shared with students (n_3 = 106) enrolled in another undergraduate-level course, CMPUT 302 Introduction to Human Computer Interaction,

which focuses on a user-centered approach to software design. Students were grouped into 21 teams and performed the cognitive walkthrough method to evaluate the usability of CBA for potential usage scenarios (i.e., actions). CBA with only one test was shared to assess usability because the other tests were not completed when teams conducted their cognitive walkthrough to reveal possible usability flaws.

A cognitive walkthrough is an analytical inspection procedure for a user interface to test and evaluate the usability issues (Atiyah et al., 2019; Shekhar & Marsden, 2018). It shows if a first-time user can understand and use the tool without any training or background knowledge (Ren et al., 2019; Shekhar & Marsden, 2018). Evaluators test different actions, and they can detect more potential problems than a user would come across in a single experience. Thus, this method helps to identify user experience issues and then take action to address these issues (e.g., Ren et al., 2019; Shekhar & Marsden, 2018). Teams performed the cognitive walkthrough method: (1) try to produce a goal, (2) search for actions available, (3) select a suitable action to progress, and (4) perform the selected action and evaluate if the progress has been made toward the initial goal (Lewis & Rieman, 2011). Each team prepared a report—a total of 21 reports—including what they were able to do and not able to do for the actions they attempted. Unfortunately, it was not possible to fix the potential usability problems of CBA as the team reports were received after data completion.

Reports were analyzed inductively from a particular to a more general perspective: from codes to themes. Analysis was conducted without searching for any specific confirmation for the results from conversation or survey data of EDPY 303 students. Reports were examined to determine which topics (i.e., usability indicators and issues) were discussed. The first step was to write memos (e.g., short statements) while reading the reports. Memos were helpful for

comparing reports with each other and making connections, and they were used to create initial codes from the reports (Creswell, 2013; Holton, 2010). Codes were simple, clear, and short and were created with the goal of representing each element of the data (Holton, 2010). Two types of coding were used: topic coding (Richards, 2009) and the constant comparative method (Charmaz, 2006). First, topic coding was conducted because it requires little understanding and interpretation (Richards, 2009). Second, following the constant comparative method, statements within the same report and then in different reports were compared to identify similarities and differences. With these two approaches, codes were created.

After the coding process, similarities between codes and their relationship with each other were investigated. Then, to reduce the number of codes, related codes were classified and combined, and themes that showed different aspects of initial codes were created (Creswell, 2013). The themes were: planned actions for CBA, unplanned actions for CBA, and actions for the assessment.

Chapter 4: Results

CBA with constructed-response and selected-response tests were designed and opened for two sections, 409 (n_1 = 290 and n_2 = 119), and one section of the EDPY 303 course, 119 students, respectively. However, the unique total number of students for selected-response tests is 98 and for constructed-response tests is 21 (see Table 6 for the number of students in each test by the course section). Thus, even though CBA was made available for a large group of students, a small group of students took CBA. A possible reason could be that the course instructors recommended CBA as an optional component of the course. In addition, both sections of the course were online making the communication between the instructors and students somewhat limited and students focusing on more grade-related and required components of the course. Among the participating students, the unique number of students who completed the survey is 61 (see Table 4 for the number of students in each survey and Table 2 for background information). In addition to the results from the assessment and survey data from participating students enrolled in EDPY 303 course, there were 21 cognitive walkthrough reports written by 106 students from CMPUT 302 course. Thus, this chapter presents the key findings from the conversation, survey, and cognitive walkthrough data. The results chapter is split into three sections: performance of CBA, student attitudes toward CBA, and cognitive walkthrough of CBA.

Table 6

Number of Students in CBA by Each Test and Section

	Total	Section 1	Section 2
Selected-response test 1	67	51	16
Selected-response test 2	77	61	16
Selected-response test 3	58	42	16
Constructed-response test 1	19	0	19
Constructed-response test 2	7	0	7

Research Question 1: Performance of CBA

CBA with Constructed-response Items

Appendix A shows the items and expected correct and incorrect responses for each item as well as item parameters for each test in CBA with the constructed-response format. Figure 8 shows an example of a constructed-response item, support, and feedback. Precision, recall, and F1-measure were calculated for each intent to evaluate CBA in intent classification (see Table 7). The recall for each constructed-response item was 100%, meaning that CBA correctly identified correct responses as correct. However, the precision measures of constructed-response items range from 80% to 100%, and F1-measure values range from 89% to 100%. That is, there are misclassifications of incorrect responses by CBA. Some of the incorrect responses to items 2 and 3 in the first test and item 4 in the second test were not interpreted accurately by CBA.

Table 7

Classification Performance for Each Item in CBA with Constructed-Response Items

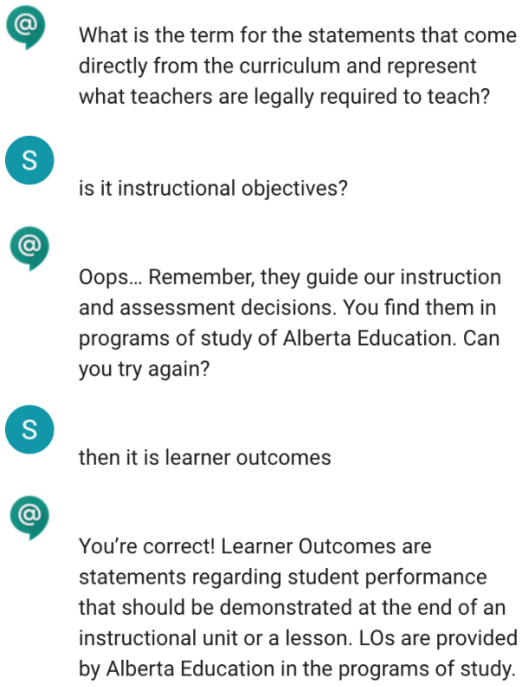
Test	Item	Precision	Recall	F1-measure
1	1	100	100	100
	2	86	100	92
	3	95	100	97
2	1	100	100	100
	2	100	100	100
	3	100	100	100
	4	80	100	89

Regarding item 2—What is the main purpose of summative assessment?—in the first test, one student's response was "fairness" and thus incorrect. However, CBA interpreted this response as correct. A possible reason is that expected correct responses for this item include the word "fair" in the trained data (see Appendix A) and thus, CBA matched the response with an inaccurate conversation path. Another student's response to item 2 was "summative assessment is assessment of learning". This response is an incorrect response to item 2; however, CBA

matched this response with an inaccurate conversation path. In detail, expected responses for item 2 include “assessment of learning” and expected responses for item 3 include “summative assessment”. Because student response includes these two parts from two expected response lists, CBA calculated similar confidence scores and chose the wrong path.

Figure 8

An Example of Constructed-Response Item from CBA



For item 3—You want to assess whether students are learning topics during the instruction. What type of assessment can you use to monitor student progress?—in the first test, one student typed “to give students a fair opportunity to demonstrate their achievement of program expectation”. However, this wording was very close to one of the expected correct responses listed for item 2 (see Appendix A), and thus CBA matched this response with an inaccurate conversation path. For this response, it should be noted that it is not clear if the student typed this response to item 3 or if they submitted another response to item 2 because their response to item 2 was already correct and received positive content feedback.

Finally, for item 4—Imagine that each year, Hogwarts, school of witchcraft and wizardry, administers a test to select talented witches and wizards. Each year students who pass a specific threshold (who got 80 out of 100) are accepted to Hogwarts. What type of grading method does Hogwarts use?—in the second test, one student's responses were "percent-based assessment" and "raw scores". However, CBA matched these responses with inaccurate conversation paths possibly because of the overlaps between the student response and the expected responses for item 3 and item 1, respectively.

It should be noted that when students were directed to an inaccurate conversation path by CBA, they attempted to answer the last item until they received an accurate response by CBA (i.e., negative or positive content feedback related to that item). Among those who received an inaccurate response by CBA and continued to send other responses, except for one of the students, they were on the accurate conversation path after one more attempt. However, one student attempted three more responses to be directed to the accurate conversation path—"percent-based assessment", "raw scores", and "absolute gradin", respectively.

In addition to intent classification, median confidence scores for correctly identified intents were reported. The median confidence scores for each correctly classified response range from 0.30 to 0.99 for correct responses and range from 0.59 to 0.98 for incorrect responses (Table 8). One interesting note from these results, all classifications of student responses to item 2 from the second test—Who designs instructional objectives?—are correct (Table 7, accuracy measures of 100%). However, the median confidence score is 0.30 even though most students typed the exact same word from the expected response list "teachers". Thus, Rasa NLU produced a lower confidence score but still correctly classified all student responses to item 2.

Table 8

Confidence Score for Each Item in CBA with Constructed-Response Items

Test	Item	Response	Confidence score
1	1	correct	0.98
		incorrect	0.93
	2	correct	0.99
		incorrect	0.59
	3	correct	0.95
		incorrect	0.98
2	1	correct	0.96
		incorrect	0.93
	2	correct	0.30
		incorrect	*
	3	correct	0.85
		incorrect	0.93
	4	correct	0.42
		incorrect	0.66

Note. * All student responses to item 2 are correct.

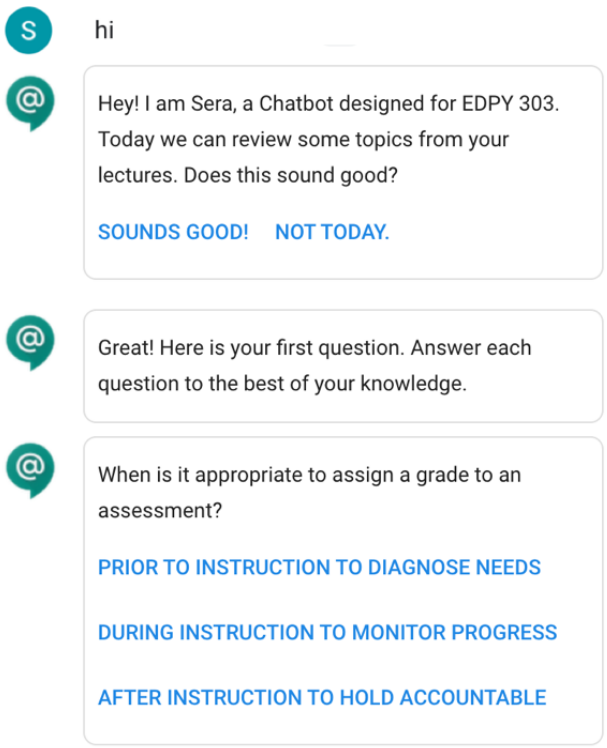
CBA showed similar performance in interpreting the students' very short (e.g., "formative") or relatively long responses (e.g., "provide a grade, assessment of learning, see if student has learned the content"). That is, it cannot be concluded that CBA was more successful in interpreting shorter or longer responses. CBA was able to interpret expressions that convey similar meaning in the responses and to direct the conversation to the accurate path (e.g., "students demonstrate achievement of learning", "to assign a grade to students"). These results can be explained by the data generation with course instructors who had the most relevant information and correct structure of this information to the CBA script. CBA was also able to interpret responses with a few misspellings accurately (e.g., "formatibe assessment"). In summary, according to the conversation data from constructed-response tests, CBA correctly analyzed student responses and moved students to the appropriate conversation path for the most part. That is, CBA showed positive content feedback if the response was correct, hints for an initially incorrect response, and negative content feedback for a second incorrect response.

CBA with Selected-Response Items

Appendix B shows the items and response options as well as item parameters and reliability for each test in CBA with the selected-response format. Figure 9 shows an example of a multiple-choice item from CBA. The precision, recall, and F1-measure of CBA were all 100%. This result was expected to be similar to CBA with constructed-response items since data was developed through conversation with course instructors with the most relevant information. This information was correctly coded into Rasa following its structure. Furthermore, each item was coded with the button options in CBA. Students only type to CBA to initiate and end a conversation, and the accuracy measures for these intents, namely *greeting* and *bye*, are also 100%. These high measures show the accuracy between the system design and implementation.

Figure 9

An Example of Selected-Response Item from CBA



In terms of the confidence score, all classifications of intents are correct. The median confidence scores for each intent were about 1, meaning that the NLU is entirely confident in classifying each input. This finding is also expected for several reasons. First, overall, NLUs produce higher confidence scores for correctly classified intents (Abdellatif et al., 2021). Second, research shows that Rasa NLU produces higher confidence scores, and the higher confidence score is highly likely to be correct classification compared to the other NLUs (Abdellatif et al., 2021). Third, similar to the reasons for high accuracy measures, CBA has produced a confidence score of 1 as it was designed with button options, and the data was structured carefully. For *greeting* and *bye*, different ways a student could greet and leave CBA were written in the CBA script, making the classifications correct and confidence scores high. These findings reveal that CBA correctly interpreted each response and moved students to the appropriate conversation path (i.e., CBA shows positive content feedback if the student response is correct while it shows negative content feedback if the response is incorrect).

Regarding unexpected inputs by students which were not defined in CBA with both constructed-response and selected-response tests (e.g., "more questions please"), CBA sent the default message "*I'm not quite sure what you mean by that response. It's okay to take your best guess to answer the question. Try again!*" The default response aimed to address the problem of unresponsiveness or inaccurate response to students. In addition, students received the default response when they typed "I don't know" or other expressions with a similar meaning (e.g., "Idk"). In the survey, students were also asked if they experienced any difficulties with CBA. Six students reported minor issues and one student reported major issues. The conversation data for those students included default responses or inaccurate conversation paths, explaining why they reported some issues with CBA.

Research Question 2: Student Attitudes toward CBA

To answer the second research question—the student attitudes toward interacting with CBA, the student responses to 12 survey items were analyzed. Nine of the items (E1 to E9) are experience-related, while the remaining three items (F1 to F3) focus on the comparison of CBA and regular assessments (Appendix D). Percentage scores of student responses are presented in Table 9, and the distribution of student responses to survey items is visualized in Figure 10.

Table 9

Percentage of Student Responses to Survey Items

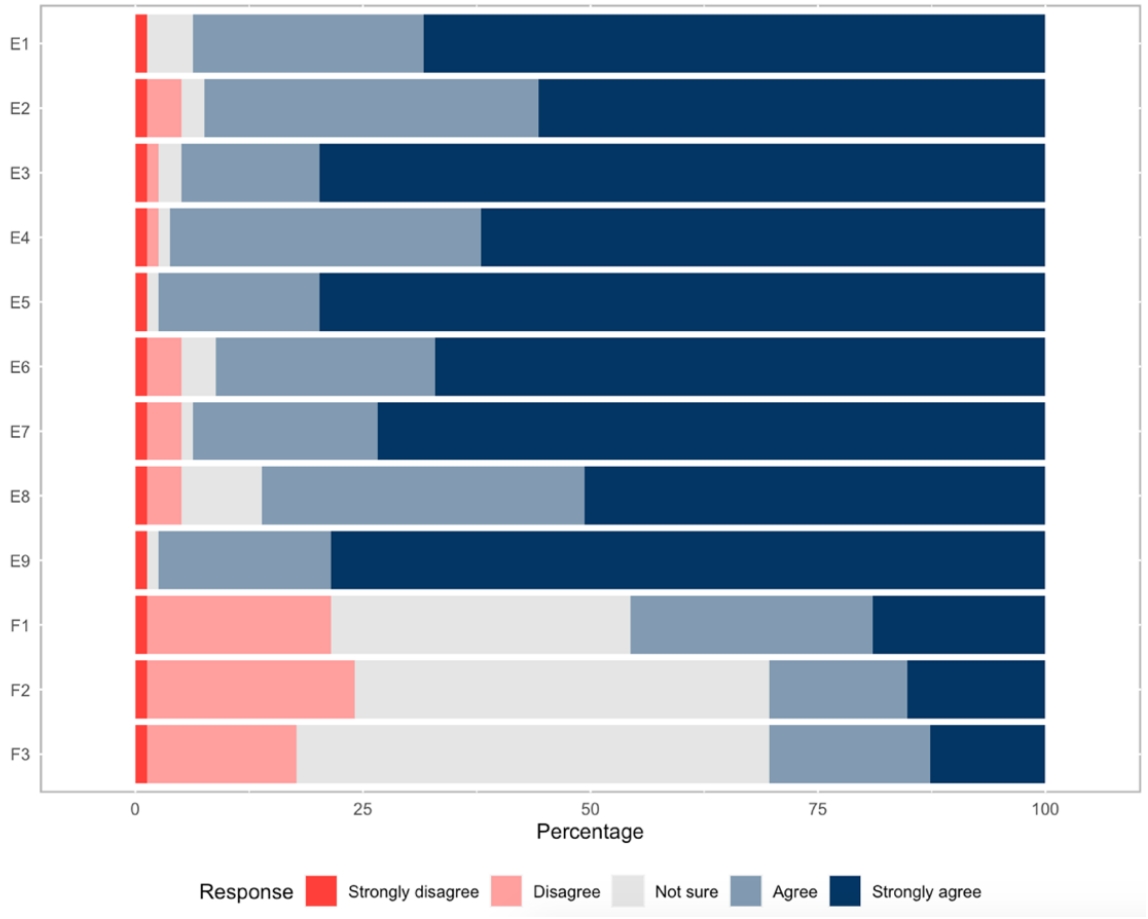
Item	Survey item	Strongly Disagree	Disagree	Agree	Strongly Agree	Not sure
E1	I found the feedback in the chatbot helpful.	1%	0%	25%	68%	5%
E2	The feedback helped me stay motivated.	1%	4%	37%	56%	3%
E3	I found the summary answer in the chatbot helpful.	1%	1%	15%	80%	3%
E4	The summary answer helped me improve my existing understanding of the concept.	1%	1%	34%	62%	1%
E5	I felt comfortable when interacting with the chatbot.	1%	0%	18%	80%	1%
E6	I was engaged during the assessment.	1%	4%	24%	67%	4%
E7	I put enough effort to answer each question.	1%	4%	20%	73%	1%
E8	The conversations in the chatbot helped me stay focused.	1%	4%	35%	51%	9%
E9	I found taking an assessment with the chatbot straightforward.	1%	0%	19%	78%	1%
F1	I prefer to take a practice exam with a chatbot compared to an online quiz.	1%	20%	27%	19%	33%
F2	I would perform better in a chatbot than in an online quiz.	1%	23%	15%	15%	46%
F3	Chatbot would provide a more accurate representation of my performance than an online quiz.	1%	16%	18%	13%	52%

Overall, students reported positive experiences with CBA and found CBA helpful and engaging. Student responses to the experience-related items were high and at similar percentages (see Table 9 for items E1 to E9). The majority of the students found the feedback (94%; item E1) and summary answer (95%; item E3) helpful and indicated that the summary answer helped them to improve their understanding (97%; item E4). This positive trend regarding student experience with content feedback suggests the importance of real-time assessment and feedback to increase the impact of intended outcomes for formative assessments on student learning, considering their use for *assessment for learning* rather than *assessment of learning*.

Students reported that they felt comfortable when interacting with the chatbot (98%; item E5) and found taking an assessment with the chatbot straightforward (98%; item E9). It should be noted that most students (72% of the students; see Table 2) did not have any previous experience using a conversational agent. Thus, this finding is promising to provide a convenient assessment environment while implementing a new form of formative assessment, CBA. In addition, even though most students reported their current technology skills as intermediate (88% of the students; see Table 2), their clear interaction with CBA shows that students are not required to have high technology skills to take an assessment with CBA. Regarding this, it should be considered that in this study, only higher education students experienced CBA. Thus, there can be an expectation that higher education students have a certain level of technology skills even though the majority reported intermediate. For very lower-grade levels (e.g., elementary or middle school students), it is possible to report some problems related to the technology skills of students.

Figure 10

Distribution of Student Responses to Survey Items



The majority of the students indicated that they were engaged during the assessment (91%; item E6) and CBA was helpful for them to stay focused (86%; item E8). To support the findings from the survey data, it is important to note that students could exit CBA before completing their attempts; however, all students finished their assessments. In addition, students continued to take other CBA tests once they were available after their first CBA experience, even though CBA was provided as an optional formative assessment tool. This also supports their positive reactions found in the survey data.

Despite their positive reactions to CBA, their responses to comparing CBA with regular assessments (e.g., online quiz) varied (see Table 9 for items F1 to F3). Forty-six percent of the students said they would prefer CBA to a regular assessment (item F1). Only 30 percent of the students indicated that they would perform better (item F2) and would be more accurately evaluated using CBA compared to a regular assessment (31%; item F3). About half of the students showed neutral reactions to this comparison. Even though this finding seems contradictory to the findings from experiences-related questions, it can be reasonable because most students (72% of the students; see Table 2) did not have any previous experience using a conversational agent. As a result, their responses to the comparison between a regular assessment and a new form of assessment were mostly neutral. On the one hand, this can be interpreted as promising not to replace regular formative assessments but to support their intended outcomes on student learning through the turn-taking conversation environment of CBA. On the other hand, this finding calls for further investigation to make a more comprehensive comparison (e.g., survey data collected for both CBA and regular assessments).

Research Question 3: Cognitive Walkthrough of CBA

Each team prepared a report—a total of 21 reports—from their cognitive walkthrough, including what they were able to do and not able to do for the actions they attempted (see Table 10). The findings of the cognitive walkthrough to assess the usability of CBA are split into three sections: planned actions for CBA (i.e., actions that CBA was trained for and not related to answering questions), unplanned actions for CBA (i.e., actions that CBA was not trained for), and actions for the assessment (i.e., actions related to answering questions).

Table 10

Actions in Cognitive Walkthrough to Evaluate Usability of CBA

Action	Number of teams
Search and open CBA in Google Chat	6
Greet the agent (e.g., type hi)	5
Affirm or deny taking CBA	3
Answer questions	21
Click an option from a previously answered question	2
Type an answer	6
Self-assessment	5
End CBA (e.g., type bye)	4
Skip a question	1
Type a single or special character or send emoji	1
Type help	1
Leave CBA without completing the assessment	1

Planned Actions for CBA

In terms of the actions of *search and open CBA in Google Chat*, *greet the agent*, *affirm or deny taking CBA*, *self-assessment*, and *end CBA,* most teams reported the usability of CBA. However, some teams reported a bug when they opened CBA: CBA sent them a question before they greeted the agent. Fortunately, teaching assistants, course instructors, and students did not encounter this bug when CBA was shared with students in EDPY 303 course. Teams also reported some suggestions to improve the usability of CBA regarding these actions. They suggested an extended introduction about CBA by the agent in terms of the acceptable actions from the users, including explanations about purpose and use (the current introduction message: *Hey! I am Sera, a Chatbot designed for EDPY 303*.) They also suggested more social interaction at the beginning before the agent asks if students want to take an assessment or not (*Today, we can review some topics from your lectures. Does this sound good?*) Even though their concern is understandable, the decision regarding short social interaction was made based on the literature (e.g., the trade-off between engagement and efficiency; Ruan et al., 2019) and the discussions with the course instructors. Also, the given explanation by the agent was short because a short

video—about two minutes long—was shared with students explaining CBA, its purpose, and its use before students' first interaction with CBA. However, their suggestion for an extended introduction is still valuable considering the concerns the teams reported (e.g., how to answer questions: type or click). Even though a video was shared with students, it is possible that some students did not watch the entire video or pay attention to the explanations. In addition to an extended introduction by CBA, one team also suggested a more distinct goodbye message (e.g., a goodbye image) by the agent as it was unclear to them if they had completed the assessment.

Unplanned Actions for CBA

Some teams attempted to perform actions that CBA was not designed for: *click an option from a previously answered question*, *skip a question*, *type a single or special character or send emoji*, *type help*, and *leave CBA without completing the assessment*. The agent failed to follow these actions (i.e., default response or inaccurate conversation path) as these actions were not aimed when building CBA.

In terms of the action of *clicking an option from a previously answered question,* this usability problem seems to be an important concern because it is possible that a student can attempt this action even though this did not occur when the EDPY 303 students interacted with CBA. To handle this issue, teams suggested making the options of the previous questions unclickable once a student chooses their answer. For the action of *skipping a question*, one team attempted to skip a question by typing "skip" in the chat. However, the agent sent the default response, as this action was not included in the CBA script. Even though the agent informed students to answer each question with the message "*Great! Here is your first question. Answer each question to the best of your knowledge.*" before sending the first question, the team still tried to skip a question. Their suggestion was to add a "skip" button for questions. Even though a

solution to this concern could be a "skip" button, the solution also depends on the purpose of the assessment and the instructor. For example, CBA was designed as an optional formative assessment, and the instructors did not suggest skipping a question while developing the questions. Thus, depending on the goals, the solution can be a "skip" button or a more detailed and clear explanation at the beginning to make sure students will not try to skip an item or a message encouraging them to answer the question.

One team also attempted to *type a single or special character or send emoji,* and they received the default response. Even though the team identified this issue as a usability problem, this is not necessarily considered a problem for CBA because CBA is designed for assessment, and it controls the direction of the conversation by asking questions and sending feedback to student responses. Thus, it is expected students should receive a default response if they type a single or special character that are not related to the content and send an emoji. In terms of the actions of *type help* and *leave CBA without completing the assessment*, one team attempted to type "help" and "leave" the assessment and suggested adding buttons for "help", "end assessment", and "restart assessment". These suggestions should be considered for further improvement of CBA because even though participating students did not attempt these actions, it is possible that students might attempt this action in the future.

Actions for the Assessment

Because all teams attempted the same action of *answering questions*, detailed information about their cognitive walkthrough is provided in Table 11. In general, most teams found the action is clear to users due to several reasons: (1) a common known format (i.e., selected-response items), (2) clickable options with blue color, larger font size, and full capitalized letters, (3) a red dot indicating unread message, (4) a loading animation once student types or selects an

answer. However, they also reported several usability problems: it is not clear (1) how to respond to the questions: type or click, (2) the total number of items on CBA, and (3) performance.

Table 11

Usability Indicators and Problems for the Action of Answering Items

Cognitive walkthrough	Usability indicators	Usability problems
Will the users try to achieve the right effect (state)?	- questions with options	- how to respond: type or click - number of items on the test
Will the user notice that the correct action is available?	- clickable button options - the standard user interface of clickable items - blue color - larger font size - full capitalized letters - red dot indicates an unread message	
Will the user associate the correct action with the effect (state) trying to be achieved?	- red dot disappears after reading - option turns blue - loading animation while a response is written	
If the correct action is performed, will the user see that progress is being made toward the solution of the task?	- feedback - next question	- typing triggers default response

For the first concern, when they typed their selected response for a multiple-choice item, the agent directed them to the correct path. However, when they typed their selected response for a true or false item, the agent either sent the default response or directed them to the wrong path. This unforeseen problem was the expectation that users would select their answers, not type. As a result of this expectation, the questions with the same response options (e.g., three true or false items in selected-response test 1) were not distinguished from each other in the CBA script. It should be noted that this problem did not occur in data collection from EDPY 303 students because they simply selected their responses from the given options instead of typing. However, the concern is important for the future use of CBA. Thus, CBA should be updated following their

suggestions: (1) explanation at the beginning about how to respond to questions; (2) make the text box unavailable for the selected-response items or (3) update the CBA script to make sure students will be directed to the correct conversation path if they type.

The second usability problem for the action of *answering questions* was the lack of information about the total number of the items on the test and the question number they were answering. CBA should be updated to provide information about how many items students will answer in CBA. Also, even though the agent says "FIRST, NEXT or FINAL QUESTION" to direct students in the assessment, CBA should be updated by numbering the questions. Teams also suggested a progress bar indicating how many questions students have answered.

The final concern was the lack of information about their performance, and the teams recommended a score bar showing how many questions they answered correctly and incorrectly. Even though this concern is reasonable from the user perspective, the goal of CBA is to provide an interactive environment for students to assess their knowledge and also scaffold their learning. Their concern and suggestion could be taken into account depending on the purpose of CBA and instructors.

In addition to the cognitive walkthrough of the actions in CBA, teams also reported valuable suggestions in general to improve the user interaction with CBA. One team suggested giving students more time to read the feedback or adding a follow-up question to confirm if users read and understood the feedback before sending the next question. The issue of feedback literacy has been discussed in the literature and suggested that the agent can enhance feedback literacy by asking questions about student understanding of the feedback. However, following the previous researchers discussing the negative impact of excessive interaction (e.g., Katz et al., 2021), a follow-up question for feedback can be judged as protracted by students. Thus, even

though their suggestion is valuable, this issue should be further investigated. The other suggestion was to send a reminder message to users if they do not respond for some time without ending the assessment. This action of CBA would help make it more interactive and human-like. CBA can be updated by scheduling a reminder to be executed after a certain time if the user stops interaction without completing the assessment. To conclude this section, the cognitive walkthrough helped view CBA from the user perspective considerably and thus identified potential usability problems that students might encounter when interacting with CBA.

Chapter 5: Discussion

This study aimed (1) to design and implement a CBA with selected-response and constructed-response items for higher education students, (2) to compare its performance in interpreting student responses, and (3) to understand its potential in advancing conventional digital assessments and obtaining further information about student attitudes toward CBA and the usability of CBA. This chapter discusses the answers to research questions and reflects on the relevant literature. It argues the practical implications of CBA, examines the study's limitations, and offers recommendations to contribute to future research on CBA.

Reflection on Research Questions

Functionality of CBA in Interpreting Student Responses

Regarding the first research question, it was aimed to investigate how accurate CBA is in interpreting and answering student responses. For the most part, CBA could consistently match students' responses and send them to an accurate conversation path. In particular, CBA with selected-response items was found accurate in interpreting all student responses. In CBA with selected-response tests, the precision, recall, and F1-measure were all 100% (match between CBA design and implementation), while they ranged from 80% to 100% in CBA with constructed-response tests (performance of CBA in understanding written responses). Regarding confidence scores of correctly classified responses, their median confidence scores were about 1 for CBA with selected-response tests, while they ranged from 0.30 to 0.99 for correct responses and from 0.59 to 0.98 for incorrect responses. These findings support the literature to a certain extent indicating that conversational agents perform well with selected responses (Huang et al., 2019; Valério et al., 2018). However, the current study showed that although CBA with selected-

response items performs better than CBA with constructed-response items, high accuracy rates of both formats show promise for measuring student knowledge and skills.

It is worth noting that CBA was always accurate in its interpretation of correct responses and sent students to the correct conversation paths. Regarding CBA with constructed-response items, it could recognize responses with a similar meaning or responses with a few misspellings. However, there were responses or other inputs (e.g., content-related or unrelated questions) that were not recognized by CBA, and the students received an inaccurate or default response. Thus, even though multiple-choice, true or false, and short response items were written in CBA to overcome the challenges of irrelevant responses (e.g., D'Mello & Graesser, 2013; Graesser, 2016) and inaccurate feedback (e.g., Lopez et al., 2021), students still experienced problems with CBA. These results back up the literature regarding the challenges in the design of CBA (e.g., Jackson & Zapata-Rivera, 2015; Yu et al., 2017). Similar to the results from previous CBA studies in the literature, the current study showed that CBA was more accurate in interpreting student responses with flexible words compared to specific words (e.g., item 2 in constructed-response test 1) (Lopez et al., 2021). However, there was no clear difference between the performance of CBA in interpreting shorter or longer responses (e.g., Huang et al., 2019; Lopez et al., 2021; Valério et al., 2018).

Student Attitudes Toward Taking an Assessment with CBA

The second question aimed to understand how students perceived their interaction with CBA. Overall, despite the reported problems regarding the functionality of CBA with constructed-response format, students indicated their positive experiences with CBA, and this result is promising to continue to work on improving CBA to become an integral part of educational assessment. The results from the survey supported the literature to a certain extent

regarding the impact of conversational agents on learning (e.g., Jackson et al., 2018; Ruan et al., 2019), motivation (e.g., Heffernan, 2003; Hong et al., 2014), and feedback (e.g., Lopez et al., 2021). Students found CBA engaging (91% item E6 and 86% item E8) and helpful (94% item E1, 95% item 3, and 97% item 4) in improving their understanding through feedback and summary answers. These results are aligned with the previous studies reporting that students are willing to take an assessment through interaction with an agent (Lopez et al., 2021; Ruan et al., 2019). However, different from the findings in the previous research (e.g., Ruan et al., 2019), most students did not indicate their choices to take assessments in CBA compared to regular assessments. Instead, most students showed neutral reactions to this comparison (about 50% for items F1 to F3). Even though this finding does not support the literature, neutral reactions to the comparison between a new form of assessment (i.e., CBA) and a regular assessment (e.g., online quiz) might be reasonable, given that this was the first interaction for the majority of the students with a conversational agent. In addition, CBA was shared with a large group of students however the participation rate was very low. On the one hand, one can argue that the availability of CBA, an interactive and personalized assessment environment, is not enough to motivate students to take assessments. On the other hand, CBA was shared as an optional and additional formative assessment tool, resulting in less attention to CBA as an optional component of the course. However, once students decided to take CBA, they showed consistency in their participation in each test throughout the course. Even though this consistency can be interpreted as a positive indicator, the results are limited by a small student group.

Usability Indicators and Issues of CBA

Previous research on conversational agents has mainly discussed the benefits and limitations of conversational agents in education using conversation or survey data. The

cognitive walkthrough of CBA significantly contributed to this research and the literature. With the cognitive walkthrough, several important usability indicators and problems of CBA have been reported. One crucial usability problem was the lack of detailed introduction by the agent. Even though CBA was designed for higher education students and a short video was shared with them before their first interaction with CBA, the reports from the cognitive walkthrough highlighted the importance of a detailed introduction by the agent. Considering what has been found in the conversation data, survey data, and cognitive walkthrough, a detailed introduction at the beginning can overcome potential problems later. It is challenging to design a perfect conversational agent that answers all expected and unexpected inputs of students. However, with a proper introduction explaining the agent's purpose and what the agent can and cannot do, some potential problems can be prevented proactively.

Practical Implications and Future of CBA

Due to the lack of interaction, it is challenging to develop engaging and motivating assessments. Thus, the betterment of conventional digital assessments is necessary and inevitable. Considering the high accuracy of the dialogue moves within CBA, student attitudes toward CBA, and the usability indicators of CBA, conversational agents and their implementation into assessments in higher education show promising results to measure and support student knowledge and skills in an interactive assessment environment. Practical implications and the future of CBA are discussed in this section in terms of item formats, classroom assessments, large-scale assessments, reliability and validity, scoring and security, and fairness and bias.

Item Formats

Tasks in CBA can be designed to allow students to demonstrate their knowledge, skills, and abilities, scaffold learning, and provide relevant feedback. Students can respond to selected-response items in a turn-taking format within a CBA, where assessment and feedback are combined to provide real-time assessment and feedback to students. Students can communicate their understanding in their own words by responding to constructed-response items within a CBA with adaptive conversations. CBA necessitates the construction of responses and the use of natural-language processing to deliver adaptive follow-up prompts that target specific information (e.g., Jackson et al., 2018). This CBA format can extend the benefits of regular constructed-response items by providing hints and asking follow-up questions. Compared with a standard digital assessment, students are more likely to benefit from a combination of assessment and feedback (in a selected-response format) and assessment, scaffolding, and feedback (in a constructed-response format) in the turn-taking conversation environment of CBA. Participating students' positive reactions to CBA support these potential benefits to a certain extent given the results from a small group of students. CBA has the potential to improve conventional format items because they combine the measurement and communication of various interacting skills (e.g., cognition, communication, and emotion) in a single, standardized setting (Jackson & Zapata-Rivera, 2015). Thus, the evidence acquired by CBA is fundamentally different from that gathered by conventional format items and CBA has the potential to expand what conventional format items can provide.

Classroom Assessment

Because formative assessment is an assessment for learning, CBA has the potential to enhance not only motivation but also learning through embedded real-time feedback (e.g., Lopez

et al., 2021). CBA can be used for formative assessment purposes and become a part of classroom assessments. They can be integrated into classroom assessments to monitor student learning and provide human-like personalized guidance to each student based on their performance. For example, in classroom assessments, previous research reports that teachers can ask one-on-one questions to assess student learning as an interactive approach (e.g., Chi et al., 2008; Fletcher, 2003). This human-to-human interaction can create an ideal assessment environment and provide significant insight into student knowledge; however, they take more time and effort, making them neither practical nor financially feasible to utilize with large student groups. CBA can address this issue in the classroom and build interactive opportunities for students by simulating human teachers and increasing motivation through sustained human-like communication. CBA can also be used in parallel with existing assessments in the classroom as they can inform teachers to offer further support where weaknesses are identified. Teachers could potentially benefit from the additional evidence provided by CBA as they can offer more insight into student knowledge compared to existing classroom assessments, such as which students require additional help (e.g., hint or prompt). CBA can also allow students to see what they know and where they need to study more.

In order to integrate CBA into classroom assessments, teachers need to decide whether CBA is suitable to use in their assessment practices in light of the aforementioned four characteristics in the literature review (i.e., the type of the system, the subject, the knowledge level of the students, and the sophistication of the dialogue strategies). For example, if there is a student cohort with low to medium ability levels, CBA can provide personalized support to each student, as well as build on each student's strengths, interests, and abilities to improve engaged and independent learning while monitoring their learning. However, a cohort with high ability

levels may not be satisfied with CBA as students may have to spend time on the concepts that they already know due to the CBA structure. Furthermore, it should also be noted that CBA is still a work in progress and has yet to become a standard feature of assessment because of the technology-related limitations. In the future, with the advancements in technology, CBA should also be developed with an interface that can allow teachers to adapt these systems to their classroom assessments easily. For example, they can be directed by a user-friendly interface about where and how to write questions, expected correct and incorrect response lists, feedback, and other features that they want in their interactive assessment environment. Thus, they should be able to write their CBA and share it with their students without help from a conversation or chatbot designer.

Large-Scale Assessments

This study suggests that CBA can be feasible for classroom assessments. However, in addition to classroom assessments, CBA may have the potential to be utilized in large-scale formative assessments to motivate students to participate in such assessments and yield more accurate results. For example, large-scale formative assessments are used to measure student achievement and inform policymakers as the results from such assessments are expected to provide guidelines to shape educational reforms. However, because formative assessments do not have any personal consequence for students given their low-stakes nature, a potential lack of motivation may threaten the reliability and validity of scores (e.g., Wise & Kong, 2005). The interactive features of CBA can support such assessments to enhance student test-taking motivation, resulting in a more accurate representation of their ability levels to inform policymakers. However, it should be noted that this potential practical implication of CBA calls for further investigation in the context of large-scale assessments.

Reliability and Validity

CBA may overcome engagement and motivation issues in formative assessments (e.g., non-effortful test-taking behavior or careless responding; Barry et al., 2010; Wise & Kong, 2005; Wise et al., 2019). They can increase student motivation and keep students motivated during an assessment (e.g., Heffernan, 2003), providing a potential to minimize the threats to the reliability and validity. For example, recent studies have shown that timely interventions such as delivering a proctor notification to the computer screen for students who appear to be disengaged can keep students motivated during the test administration (Wise et al., 2019; Wise et al., 2022). This emphasizes the necessity of interactivity that is missing in regular assessments to motivate students during assessments. Although there appear to be promising approaches to maintain or increase motivation through some sort of interaction (e.g., a timely intervention), there is still a need for natural interactivity in formative assessments rather than performing a reactive intervention before students become completely disengaged from the assessment. Providing students with the opportunity in CBA to be interactive during a test administration can enhance their motivation towards the assessment process, resulting in more effortful test-taking behavior and reliable and valid test results. Survey results from this study support this possible practical implication of CBA to some extent. Low participation rates in both CBA and survey limit our interpretation to a small student group. Future research should investigate the impact of CBA on test-taking engagement and the reliability and validity of test scores obtained from CBA considering its structure (e.g., hints and feedback).

Scoring and Security

Practical implications of CBA can also be discussed in terms of test scoring and security. Automated scoring of constructed-response items was not part of the CBA in this study because

the focus was on the functionality and usability of CBA as well as student attitudes toward CBA. However, a simple automated scoring would be integrated as the constructed-response items were short responses. In addition, a more complicated AI-based automated scoring can be integrated into a CBA with essay items as it is itself an AI-based tool. In addition to automated scoring, the shift to online and remote mode has been studied from an assessment point-of-view and the discussions have brought the concept of online, remote, and recently AI-based proctoring (e.g., Motwani et al., 2021). The expectation would be that CBA has more flexibility compared to other digital assessments to integrate an AI-based proctor system as it is already an AI-based tool. It should be noted that the discussions refer to the potential implications of CBA in terms of automated scoring and test security and thus future studies should investigate these issues. Furthermore, the discussions here are not related to the advantages or disadvantages of using automated scoring or proctoring in formative assessments but rather the potential of CBA in terms of the integration of such systems.

Fairness and Bias

As with any assessment tool, it is important that CBA is fair and free from bias against any group of students. There are several issues that can lead to unfair assessments and scores including content, format, and scoring. These issues require a detailed discussion, however, in the context of this study, it would be appropriate to address the assessment format. In terms of the assessment format, even though technology skills required to take CBA may not be considered more demanding than any other digital assessments, it is likely that some student subgroups can be advantaged over others. On the other hand, non-effortful test-taking behavior occurs at different proportions across student subgroups, leading to additional inequities among them. As a result, achievement gaps between student subgroups may widen (Soland, 2018a,

2018b). Similar to the discussion on reliability and validity, an interactive assessment environment of CBA can enhance motivation towards the assessment process and thus can increase test-taking engagement, addressing one potential source of fairness.

Research on bias with respect to AI has been discussing bias-related issues (e.g., gender bias; associating "teacher" with female and "doctor" with male, or designing conversational agents to be female) and the ability to monitor and address bias (Feine et al., 2019; Makhortykh et al., 2021; Nadeem et al., 2020). The underlying reason behind these issues is that AI can learn desirable and undesirable actions or behaviors depending on the information provided to train it (e.g., Nadeem et al., 2020). This makes educational researchers and practitioners who design CBA responsible to be careful not to introduce bias to their systems. In addition, it is not realistically possible to capture all complexities of a group of students with a conversational agent. Thus, a conversational agent designed and used for many different purposes (e.g., tutoring, assessment, answering frequently asked questions) is less likely to be unbiased. With its specific use, CBA can capture more characteristics of a student group. To detect and address possible bias issues in CBA systems, they can be tested by diverse groups (e.g., cognitive walkthrough) and then shared with the target group (e.g., Nadeem et al., 2020). In addition, the issue of designing text and voice-based conversational agents to be mainly female (i.e., female voice or character) has been discussed in the literature (e.g., Brahnam & De Angeli, 2012). In this study, the computer agent in CBA used the name "Sera" which can call female gender associations. Unfortunately, this decision may cause gender stereotypes (e.g., female teachers). This was an unintentional gender bias in the design of CBA. Even though the intention was not to prefer one gender over another, this study showed a tendency to use the female gender. Future research should mitigate the gender bias in the design of CBA.

Automation bias should also be discussed as students can show this type of bias even though a CBA design is free from any bias. Automation bias refers to inappropriate decision-making as a result of overreliance or overdependence on AI-based systems (Lyell & Coiera, 2017; Parasuraman & Riley, 1997). Considering the current limitations of CBA (e.g., inaccurate conversation path), if students over-rely on CBA, they can follow inappropriate feedback. To address this issue, we should aim to design CBA with high reliability as much as possible with the help of content experts, pilot studies, and cognitive walkthroughs. Even though there is no research discussing automation bias for conversational agents in education, using content feedback rather than progress feedback can mitigate the automation bias as students are provided with the content information. Thus, in the current CBA design in this study, the use of content feedback could alleviate possible automation bias. It should be noted that even though some initial discussions around fairness and bias were attempted in the context of CBA, these issues need to be further investigated. To conclude the section on practical implications and the future of CBA, the possible uses of CBA can be extended as they have the potential to be an integral component of education and assessment in the future and to make significant improvements to educational assessments.

Limitations and Future Research

Technology-related Limitations

The design and implementation of CBA have several technology-related limitations. In this study, similar to previous research, CBA was unable to handle all students' written responses and thus failed to direct them to the accurate conversation path. As a result of these failures, students received the default response or irrelevant response, and they reported technological issues in their survey responses. Thus, despite the fact that short response items were created to

address the issues of irrelevant or inaccurate responses, students still experienced problems with CBA. Previous research has discussed that conversational agents have not yet reached the level of development needed to understand content to the extent that we believe necessary (e.g., Maedche et al., 2019). As a result, CBA is not yet sufficiently evolved to comprehend or evaluate the level of information in student responses. As technology enhances, CBA can be progressively more capable to address these issues. However, we should still be able to address the current problems with the existing technology considering the design-related issues discussed below.

Design-related Limitations

The pilot test of CBA with the course instructors and teaching assistants was an essential part of CBA design. After testing CBA, necessary changes were made. However, the changes were limited to the small group of people who tried CBA and gave feedback for further improvement. Future studies should provide close and thoughtful attention and collect more information from a larger group during the design phase of CBA. In addition, those who design CBA can ask more experts to revise and expand the list of expected responses. This will allow to include a better list of expected correct and incorrect responses and also other inputs, improving the functionality of CBA.

In addition to the pilot study of CBA, another design-related limitation of this study is due to the cognitive walkthrough. There was no study in the literature investigating conversational agents in education that used a cognitive walkthrough approach to evaluate the usability of their systems. Thus, cognitive walkthrough was an important contribution to this study and also to the literature more than expected. However, CBA with only one test was assessed in terms of its usability because the other tests were not ready to be evaluated when

teams conducted their cognitive walkthrough. Furthermore, cognitive walkthrough aims to identify and fix the potential usability problems, however, even though the potential usability problems of CBA were identified it was not possible to fix them as the reports were received after data completion. Future studies should consider a cognitive walkthrough approach for their entire system and make changes accordingly before sharing CBA with students. In addition to the pilot study, the cognitive walkthrough has the potential to make a substantial improvement in the performance of CBA.

The generalizability of the results from this study may be limited due to using only short-answer items in CBA with the constructed-response format. Further research can investigate the accuracy of CBA in interpreting student responses to essay items. The use of selected-response items in CBA can also be considered a limitation due to the restricted interaction between the test-taker and the agent. However, it was aimed that compared to a regular digital assessment, students taking CBA with selected-response items can more likely benefit from a combination of assessment and feedback in the turn-taking conversation environment of CBA. Student responses to survey items also support this interpretation. Future research can design CBA including both selected-response and constructed-response items instead of designing separate selected-response and constructed-response tests to increase the interaction between the test-taker and the agent throughout the assessment. In addition, CBA presented the items in a non-adaptive item format with a fixed-length form. Research should study an adaptive item format with a custom test form for each student. This further research would also increase and provide further information on test security (e.g., the use of common items; Lee et al., 2019).

Finally, CBA was designed for higher education students on qualitative content. Future research can design CBA for lower-grade levels or on quantitative subjects and investigate the

functionality of CBA and students' attitudes toward CBA. CBA has the potential to play a critical role in the future of assessments for both higher education and lower-grade levels, however, technology skills (e.g., difficulty with keyboard) might be a problem for early grades, calling for further research on CBA for lower-grade levels. In addition, survey data consisted of only quantitative data limiting the understanding of student experience with CBA. For example, even though survey data indicates that seven students reported some issues with CBA, there is no further information about student feelings due to these issues (e.g., frustration). Future research should consider including qualitative survey items or designing individual or focus group interviews with students to better understand their attitudes toward CBA.

Implementation-related Limitations

One of the implementation-related limitations is having conversation data from a small group of students. Thus, researchers should carefully examine the conclusions drawn from this study. Further research with a larger sample size would enable a more thorough understanding of how CBA performs with selected and constructed-response items and how CBA can advance student assessment experiences. Future research with a larger sample size would also allow for a more thorough understanding of CBA for different subgroups of students (e.g., gender, content knowledge, technology level) and thus for fairness issues in the context of CBA.

The current work is also limited by the unbalanced student groups who took CBA with selected and constructed-response formats. Also, one section of the course took both CBA formats, while one section took only the selected-response format. Thus, this study was for exploratory research purposes and could not provide a comprehensive understanding of two CBA formats and student attitudes toward CBA. This limits the study to fully addressing the research questions even though the results present foundational support for the inclusion of

conversations in assessments. Future research can collect data from a more balanced student group and even consider experimental research to compare CBA formats or investigate the impact of CBA on students (e.g., a control group where students take only traditional assessments and an experimental group where students experience CBA).

Despite these limitations of the current work, the results from this study show the promise of CBA in measuring student knowledge and skill and enhancing their assessment experiences. Furthermore, CBA with both formats collected student response data and log entries for each action, including date, time, test-taker, and event name and type. CBA with constructed-response format also contained information about student misconceptions, which is not always possible to obtain through conventional digital assessments. This shows the potential of CBA in advancing conventional digital assessments and obtaining further information about student knowledge and behavior. Future research should collect more information through CBA (i.e., student response data and log entries) to understand and address the issues in learning and assessment (e.g., does student performance increase over time compared to traditional assessments, or do students show more effortful test-taking behavior compared traditional assessments?)

Closing Thoughts

Technological advancements continue to yield cutting-edge applications that can enhance the quality and effectiveness of educational practices. In recent years, novel technologies (e.g., conversational agents) involving AI and NLP have enabled computers to process linguistic input more accurately and create human-like communication to interact with learners. To date, conversational agents in education have been mainly used for instructional purposes (e.g., Goel & Polepeddi, 2016; Graesser et al., 1999; Howard et al., 2017). The current conversational

agents have one important impact: enhancing motivation. With this particular impact on students, they can become even more effective and be utilized for different purposes in education (e.g., CBA).

CBA can provide both interactivity and assistance, which are missing in conventional digital assessments. They can be administered to motivate students to take assessments by holding conversations with a conversational agent and thereby enhancing their assessment experiences. Richer interaction opportunities with CBA can advance personalization and improvement of learning and motivation. Through student interaction, CBA can effectively leverage content to target information (Jackson & Zapata-Rivera, 2015). They can direct students on what to do next, ask questions, provide hints, repeat or rephrase questions, hold social interactions, and provide feedback on the quality of responses. The future of education will involve more conversational agents that will guide learners across all stages of learning—teaching, assessment, and feedback. To conclude, CBA has a valuable role in the future of education and assessment as it is expected to become increasingly more common and progressively more capable as technology advances.

References

Abdellatif, A., Badran, K., Costa, D., & Shihab, E. (2021). A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. *IEEE Transactions on Software Engineering*. http://dx.doi.org/10.1109/TSE.2021.3078384

Aleven V., Ogan A., Popescu O., Torrey C., & Koedinger K. (2004). Evaluating the Effectiveness of a Tutorial Dialogue System for Self-Explanation. In J. C. Lester, R. M. Vicari & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems: Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-30139-4_42

Aleven, V., Popescu, O., & Koedinger, K. R. (2001). Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.), In *Proceedings of the 10th International Conference on Artificial Intelligence in Education*. Amsterdam: IOS Press. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.353.340&rep=rep1&type=pdf

Atiyah, A., Jusoh, S., & Alghanim, F. (2019). Evaluation of the naturalness of chatbot applications. In *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 359-365). IEEE. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8717455

Attali, Y., & Powers, D. (2008). Effect of immediate feedback and revision on psychometric properties of open-ended GRE subject test items. *ETS Research Report Series*. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02107.x

Azevedo, R., Witherspoon, A., Graesser, A., McNamara, D., Chauncey, A., Siler, E., Cai, Z., Rus, V., & Lintean, M. (2009). MetaTutor: Analyzing self-regulated learning in a

tutoring system for biology. In *Artificial Intelligence in Education*. IOS Press. http://dx.doi.org/10.3233/978-1-60750-028-5-635

Bailey, D., Southam, A., & Costley, J. (2021). Digital storytelling with chatbots: Mapping L2 participation and perception patterns. *Interactive Technology and Smart Education, 18*(1), 85-103. http://dx.doi.org/10.1108/ITSE-08-2020-0170

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Beun, R. J., de Vos, E., & Witteman, C. (2003). Embodied Conversational Agents: Effects on Memory Performance and Anthropomorphisation. In T. Rist, R. S. Aylett, D. Ballin & J. Rickel (Eds.), *Intelligent Virtual Agents: Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-39396-2_52

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, *24*(3), 139-153. https://doi.org/10.1016/j.intcom.2012.05.001

Cai, Z., Graesser, A. C., Millis, K. K., Halpern, D. F., Wallace, P. S., Moldovan, C., & Forsyth, C. (2009). ARIES: An intelligent tutoring system assisted by conversational agents. In *Artificial Intelligence in Education* (pp. 796-796). IOS Press. http://dx.doi.org/10.3233/978-1-60750-028-5-796

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis. Los Angeles, CA: SAGE.*

Chi, M. T., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*(2), 301-341. http://dx.doi.org/10.1080/03640210701863396

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz & J. Vassileva (Eds.), In *Proceedings of the 8th International Conference on User Modeling* (pp. 137-147). Springer, Berlin, Heidelberg. Retrieved from https://link.springer.com/content/pdf/10.1007%2F3-540-44566-8.pdf

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, *29*(3), 241-250. http://dx.doi.org/10.1080/1358165042000283101

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. Los Angeles, CA: SAGE.

Davis, R. O. (2018). The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review*, *24*, 193-209. https://doi.org/10.1016/j.edurev.2018.05.002

Dodds, P., & Fletcher, J. D. (2004). Opportunities for new "smart" learning environments enabled by next-generation web capabilities. *Journal of Educational Multimedia and Hypermedia*, *13*(4), 391-404. Retrieved from https://www.learntechlib.org/primary/p/6583/

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent

tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284-332. http://dx.doi.org/10.1007/s40593-014-0017-9

D'Mello, S., & Graesser, A. (2013). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, *2*(4), 1-39. http://dx.doi.org/10.1145/2395123.2395128

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems*, *22*(4), 53-61. http://dx.doi.org/10.1109/MIS.2007.79

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). Gender bias in chatbot design. In *International Workshop on Chatbot Research and Design* (pp. 79-93). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-39540-7_6

Fletcher, J. D. (2003). Evidence for learning from technology-assisted instruction. In J. H. F. O'Neil & R. Perez (Eds.), In *Technology Applications in Education: A Learning View* (pp. 79-99). Erlbaum, Hillsdale, NJ.

Goel, A. K., & Joyner, D. A. (2017). Using AI to teach AI: Lessons from an online AI class. *AI Magazine*, *38*(2), 48-59. http://dx.doi.org/10.1609/aimag.v38i2.2732

Goel, A. K., & Polepeddi, L. (2016). Jill Watson: A virtual teaching assistant for online education. *Georgia Institute of Technology*. Retrieved from http://hdl.handle.net/1853/59104

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education*, *26*(1), 124-132. http://dx.doi.org/10.1007/s40593-015-0086-4

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612-618. http://dx.doi.org/10.1109/TE.2005.856149

Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. (2007). Emotions and learning with AutoTutor. *Frontiers in Artificial Intelligence and Applications*, *158*, 569-571.

Graesser, A. C., Jackson, G. T., Matthews, E. C., Mitchell, H. H., Olney, A., Ventura, M., ... & Tutoring Research Group. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Retrieved from https://escholarship.org/uc/item/6mj3q2v1

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes*, *45*(4-5), 298-322. http://dx.doi.org/10.1080/01638530802145395

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180-192. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03195563.pdf

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, *23*(5), 374-380. http://dx.doi.org/10.1177/0963721414540680

Graesser, A. C., Moreno, K., Marineau, J., Adcock, A., Olney, A., Person, N., & Tutoring Research Group. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialogue or the talking head? In *Proceedings of Artificial Intelligence in Education* (Vol.

4754). Retrieved from https://cpb-us-w2.wpmucdn.com/blogs.memphis.edu/dist/d/2954/files/2019/10/AutoTutor-improves-deep-learning-of-computer-literacy-Is-it-the-dialog-or-the-talking-head.pdf

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104-137. http://dx.doi.org/10.3102/00028312031001104

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*, 495-522. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/acp.2350090604

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & Tutoring Research Group. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, *1*(1), 35-51. http://dx.doi.org/10.1016/S1389-0417(99)00005-4

Heffernan, N. T. (2003). Web-based evaluations showing both cognitive and motivational benefits of the Ms. Lindquist tutor. In *Artificial intelligence in Education* (pp. 115-122). IOS Press.

Holton, J. A. (2010). The coding process and its challenges. *Grounded Theory Review: An International Journal, 9*(1), 21-40.

Hong, Z. W., Chen, Y. L., & Lan, C. H. (2014). A courseware to script animated pedagogical agents in instructional material for elementary students in English education. *Computer Assisted Language Learning*, *27*(5), 379-394. https://doi.org/10.1080/09588221.2012.733712

Howard, C., Jordan, P., Di Eugenio, B., & Katz, S. (2017). Shifting the load: A peer dialogue agent that encourages its human collaborator to contribute more to problem

solving. *International Journal of Artificial Intelligence in Education*, *27*(1), 101-129. http://dx.doi.org/10.1007/s40593-015-0071-y

Huang, W., Hew, K. F., & Gonda, D. E. (2019). Designing and evaluating three chatbot-enhanced activities for a flipped graduate course. *International Journal of Mechanical Engineering and Robotics Research, 8*(5), 813-818. http://dx.doi.org/10.18178/ijmerr.8.5.813-818

Jackson, G. T., Castellano, K. E., Brockway, D., & Lehman, B. (2018). Improving the measurement of cognitive skills through automated conversations. *Journal of Research on Technology in Education*, *50*(3), 226-240. http://dx.doi.org/10.1080/15391523.2018.1452655

Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos*, *39*(60), 31-48.

Jackson, G. T., & Graesser, A. C. (2007). Content matters: An investigation of feedback categories within an ITS. In *Artificial Intelligence in Education* (pp. 127-134). IOS Press. Retrieved from https://books.google.ca/books?hl=en&lr=&id=GEK93NUHdXYC&oi=fnd&pg=PA127&dq=Content+matters:+An+investigation+of+feedback+categories+within+an+ITS&ots=Rtpdhp8E32&sig=5PF2ggLv-fL6_qwwMAUu388vvVc#v=onepage&q=Content%20matters%3A%20An%20investigation%20of%20feedback%20categories%20within%20an%20ITS&f=false

Jackson, G. T., & Zapata-Rivera, D. (2015). Conversation-based assessment. *R&D Connections*, *25*, 1-8. Retrieved from https://origin-www.ets.org/Media/Research/pdf/RD_Connections_25.pdf

Jia, J. (2003). The study of the application of a keywords-based chatbot system on the teaching of foreign languages. arXiv preprint cs/0310018. Retrieved from https://arxiv.org/abs/cs/0310018

Jordan P., Albacete P., & Katz S. (2018). A Comparison of Tutoring Strategies for Recovering from a Failed Attempt During Faded Support. In C. Penstein Rosé et al. (Eds.), *Artificial Intelligence in Education: Lecture Notes in Computer Science* (Vol. 10947). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-93843-1_16

Jordan, P., Albacete, P., & Katz, S. (2016). Exploring contingent step decomposition in a tutorial dialogue system. In *Extended Proceedings of UMAP*. Retrieved from http://ceur-ws.org/Vol-1618/LBR2.pdf

Katz, S., Albacete, P., Chounta, I. A., Jordan, P., McLaren, B. M., & Zapata-Rivera, D. (2021). Linking dialogue with student modelling to create an adaptive tutoring system for conceptual physics. *International Journal of Artificial Intelligence in Education*, 1-49. http://dx.doi.org/10.1007/s40593-020-00226-y

Kerly A., & Bull S. (2008). Children's Interactions with Inspectable and Negotiated Learner Models. In B. P. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), *Intelligent Tutoring Systems: Lecture Notes in Computer Science* (Vol. 5091). Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-69132-7_18

Kerly, A., Ellis, R., & Bull, S. (2008a). Conversational agents in E-Learning. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 169-182). Springer, London. Retrieved from https://link.springer.com/chapter/10.1007/978-1-84882-215-3_13

Kerly A., Ellis R., & Bull S. (2008b). CALMsystem: A Conversational Agent for Learner Modelling. In: Ellis R., Allen T., Petridis M. (Eds.), *Applications and Innovations in Intelligent Systems*. SGAI 2007. Springer, London. http://dx.doi.org/10.1007/978-1-84800-086-5_7

Kopp, K. J., Britt, M. A., Millis, K., & Graesser, A. C. (2012). Improving the efficiency of dialogue in tutoring. *Learning and Instruction*, *22*(5), 320–330. http://dx.doi.org/10.1016/j.learninstruc.2011.12.002

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings IEEE international conference on Advanced Learning Technologies* (pp. 43-46). IEEE. http://dx.doi.org/10.1109/ICALT.2001.943850

Lee, Y. H., Haberman, S. J., & Dorans, N. J. (2019). Use of adjustment by minimum discriminant information in linking constructed-response test scores in the absence of common items. *Journal of Educational Measurement*, *56*(2), 452-472. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1111/jedm.12216

Lewis, C., & Rieman, J. (2011). *Task-centered user interface design*. https://hcibib.org/tcuid/tcuid.pdf

Lopez, A. A., Guzman-Orth, D., Zapata-Rivera, D., Forsyth, C. M., & Luce, C. (2021). Examining the Accuracy of a Conversation-Based Assessment in Interpreting English Learners' Written Responses. *ETS Research Report Series*. http://dx.doi.org/10.1002/ets2.12315

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based digital assistants: opportunities, threats, and research perspectives. *Business & Information Systems Engineering*, *61*(4), 1-29. http://dx.doi.org/10.1007/s12599-019-00600-8

Makhortykh, M., Urman, A., & Ulloa, R. (2021). Detecting race and gender bias in visual representation of AI on web search engines. In *International Workshop on Algorithmic Bias in Search and Recommendation* (pp. 36-50). Springer, Cham. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-78818-6_5

McAndrew, P., & Scanlon, E. (2013). Open learning at a distance: lessons for struggling MOOCs. *Science*, 342(6165), 1450-1451. http://dx.doi.org/10.1126/science.1239686

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222-233. http://dx.doi.org/10.3758/BF03195567

Motwani, S., Nagpal, C., Motwani, M., Nagdev, N., & Yeole, A. (2021). AI-based proctoring system for online tests. In *Proceedings of the 4th International Conference on Advances in Science & Technology*. http://dx.doi.org/10.2139/ssrn.3866446

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. *ACIS 2020 proceedings*. http://hdl.handle.net/10453/146564

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, *24*, 427-469. http://dx.doi.org/10.1007/s40593-014-0029-5

Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., & Cassell, J. (2012). "Oh dear Stacy!": Social interaction, elaboration, and learning with teachable agents. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 39–48. https://doi.org/10.1145/2207676.2207684

Ogan, A., Finkelstein, S., Walker, E., Carlson, R., & Cassell, J. (2012). Rudeness and rapport: Insults and learning gains in peer tutoring. *Intelligent Tutoring Systems*, 11–21. https://doi.org/10.1007/978-3-642-30950-2_2

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Rafla, A., & Kennington, C. (2019). Incrementalizing RASA's Open-Source Natural Language Understanding Pipeline. arXiv preprint arXiv:1907.05403.

Ren, R., Castro, J. W., Acuña, S. T., & de Lara, J. (2019). Evaluation techniques for chatbot usability: A systematic mapping study. *International Journal of Software Engineering and Knowledge Engineering*, *29*(11n12), 1673–1702. https://doi.org/10.1142/S0218194019400163

Richards, L. (2009). *Handling qualitative data: A practical guide. Los Angeles, CA: SAGE.*

Rozin, P., & Cohen, A. B. (2003). Reply to commentaries: Confusion infusions, suggestives, correctives, and other medicines. *Emotion*, *3*(1), 92-96. http://dx.doi.org/10.1037/1528-3542.3.1.92

Ruan, S., Jiang, L., Xu, J., Tham, B. J. K., Qiu, Z., Zhu, Y., Murnane, E. L., Brunskill, E., & Landay, J. A. (2019). Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). http://dx.doi.org/10.1145/3290605.3300587

Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in conversational intelligent tutoring systems. *AI Magazine*, *34*(3), 42-54. http://dx.doi.org/10.1609/aimag.v34i3.2485

Rus, V., Niraula, N. B., & Banjade, R. (2015). DeepTutor: An effective, online intelligent tutoring system that promotes deep learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (pp. 4294-4295). AAAI Press. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10019/9857

Shahriar Khan, F., Al Mushabbir, M., Sabik Irbaz, M., & Abdullah Al Nasim, M. D. (2021). End-to-end natural language understanding pipeline for Bangla conversational agents. arXiv e-prints arXiv:2107.05541.

Shekhar, A., & Marsden, N. (2018). Cognitive walkthrough of a learning management system with gendered personas. In *Proceedings of the 4th Conference on Gender & IT* (pp. 191-198). https://doi.org/10.1145/3196839.3196869

Snowman, J., & McCown, R. (2015). *Psychology applied to teaching*. Belmont, CA: Wadsworth Cengage Learning.

So, Y., Zapata-Rivera, D., Cho, Y., Luce, C., & Battistini, L. (2015). Using trialogues to measure English language skills. *Educational Technology & Society*, *18*(2), 21-32. Retrieved from https://www.jstor.org/stable/pdf/jeductechsoci.18.2.21.pdf?casa_token=WtQiHLueZ5AAAAAA:zkjwNGUWv8NsTazVwcT3xkcwijTl8kQYgrcai-

xZtQaXp8MSwOyNDDzkJR0zkqeSbY5Q32yBBsC1XFyhOcwXXO5YEFbuQDpMy-HORMAVylKdrsWHIN1V

Soland, J. (2018a). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, *120*(12), 1–26.

Soland, J. (2018b). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Candello, H. (2018). Chatbots explain themselves: Designers' strategies for conveying chatbot features to users. *Journal on Interactive Systems, 9*(3), 61-79.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. http://dx.doi.org/10.1080/00461520.2011.611369

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3-62. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1080/03640210709336984

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wang, W. Y., Finkelstein, S., Ogan, A., Black, A. W., & Cassell, J. (2012). "Love ya, jerkface": Using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 20–29. http://dl.acm.org/citation.cfm?id=2392800.2392805

Weerasinghe, A., Mitrovic, A., & Martin, B. (2009). Towards individualized dialogue support for ill-defined domains. *International Journal of Artificial Intelligence in Education*, *19*(4), 357-379.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. Retrieved from https://dl.acm.org/doi/pdf/10.1145/365153.365168?casa_token=DnTo5s249mEAAAAA:q9AxHLU9PPnIJSpvKkQp0GfVbbL2Ns9Y89TdY7T9g_R4ARGSNHPT1a15NXNo23RchJqmaa8nT6VIPw

Windiatmoko, Y., Hidayatullah, A. F., & Rahmadi, R. (2020). Developing Facebook chatbot based on deep learning using RASA framework for university enquiries. arXiv preprint arXiv:2009.12341. http://dx.doi.org/10.1088/1757-899X/1077/1/012060

Winkler, R., Söllner, M., Neuweiler, M. L., Conti Rossini, F., & Leimeister, J. M. (2019). Alexa, can you help us solve this problem? How conversations with smart personal assistant tutors increase task group outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6). http://dx.doi.org/10.1145/3290607.3313090

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., Kuhfeld, M. R., & Cronin, J. (2022) Assessment in the time of COVID-19: Understanding patterns of student disengagement during remote low-stakes testing. Educational Assessment, 1-16. https://doi.org/10.1080/10627197.2022.2087621

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183-192. https://doi.org/10.1080/08957347.2019.1577248

Wisher, R. A., & Fletcher, J. D. (2004). The case for advanced distributed learning. *Information and Security*, *14*, 17-25. Retrieved from https://it4sec.org/bg/system/files/14.01_Wisher_Fletcher.pdf

Yang, H. C., & Zapata-Rivera, D. (2010). Interlanguage pragmatics with a pedagogical agent: The request game. *Computer Assisted Language Learning*, *23*(5), 395-412. http://dx.doi.org/10.1080/09588221.2010.520274

Yu, H., Miao, C., Leung, C., & White, T. J. (2017). Towards AI-powered personalization in MOOC learning. *npj Science of Learning*, *2*(15), 1-5. http://dx.doi.org/10.1038/s41539-017-0016-3

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3-17. https://doi.org/10.1207/s15326985ep2501_2

Appendix A

CBA with Constructed-Response Tests

Table A1

Items in CBA with the Constructed-Response Test 1

Item	Expected correct responses	Expected incorrect responses
1. What is the term for the statements that come directly from the curriculum and represent what teachers are legally required to teach?	- learner outcomes - learning outcomes - LO	- Instructional objective - behavior outcomes - performance outcomes - learning target - IO
2. What is the main purpose of summative assessment?	- measuring learning - evaluating learning - assessing knowledge - evaluate student proficiency - assigning grades - reporting learning - give students a fair chance to show what they have learned	- Supporting learning - Adjusting teaching strategies - informing instruction - assessment of learning
3. You want to assess whether students are learning topics during the instruction. What type of assessment can you use to monitor student progress?	- Formative assessment - Assessment for learning	- Summative assessment - constructed-response - selected-response - multiple-choice - matching - extended constructed-response - portfolio assessment - performance-based assessment - short answer item

Table A2

Items in CBA with the Constructed-Response Test 2

Item	Expected correct responses	Expected incorrect responses
1. What information does a holistic rubric provide about student performance?	- General feedback - overall impression - a single judgment	- Separate judgment - specific feedback - Detailed feedback - separate feedback for scoring categories - specific
2. Who designs instructional objectives?	- Teachers - classroom teachers	- Alberta education - Province - School board - school principals
3. Identify the question type in the example. EXAMPLE: You developed a math question for your students as follows: The three interior angles of a triangle will always have a sum of _____.	- Restricted constructed-response - fill-in the blank - completion item - completion	- Summative assessment - formative assessment - constructed-response - selected-response - multiple-choice - matching - extended constructed-response - portfolio assessment - performance-based assessment - short answer item
4. Imagine that each year, Hogwarts, school of witchcraft and wizardry, administers a test to select talented witches and wizards. Each year students who pass a specific threshold (who got 80 out of 100) are accepted to Hogwarts. What type of grading method does Hogwarts use?	- Criterion-referenced assessment - Absolute	- Norm-referenced assessment - Relative

Table A3

Item Parameters for CBA with Constructed-Response Tests

	CR test 1		CR test 2	
	b	*a*	*b*	*a*
Item 1	0.74	0.56	0.71	0.67
Item 2	0.74	0.35	1	0
Item 3	0.89	0.23	0.43	0.75
Item 4	—	—	0.86	0.34

Note. CR: constructed-response; *b*: item difficulty parameter; *a*: item discrimination parameter. Reliability coefficients are not reported due to the small sample sizes.

Appendix B

CBA with Selected-Response Tests

Table B1

Items in CBA with the Selected-Response Test 1

Item	Response options
1. When is it appropriate to assign a grade to an assessment?	- Prior to instruction to diagnose needs - During instruction to monitor progress - After instruction to hold accountable
2. Which of these pairs of words BEST describes key characteristics of an instructional objective?	- challenging and desirable - observable and measurable - limited and precise
3. Who is responsible for developing instructional objectives?	- Classroom teacher - School board - Alberta Education
4. Planning formative assessment requires numerous steps such as identifying objectives, learning and assessment activities, and topics for instruction.	- True - False
5. The responses that students give on summative assessments should be related to	- the objectives and instructions provided - things all students have learned to do - skills all teachers could emphasize
6. Summative assessment is used to monitor students' progress in terms of learning.	- True - False
7. Which of the following is the MOST important preparation for summative assessments?	- Providing students with chapter reviews - Providing students with good instruction - Teaching students test-wiseness
8. One of the main roles of teachers during formative assessment processes is to make fair assessments.	- True - False

Table B2

Items in CBA with the Selected-Response Test 2

Item	Response options
1. Guessing is MORE likely to occur in constructed-response (CR) items than selected (SR) items.	- True - False
2. One property of a high-quality restricted constructed-response (CR) item is that it	- allows freedom of expression - uses guiding words to clarify expectations - eliminates objectivity in scoring
3. A holistic rubric has separate descriptors for each criterion.	- True - False
4. Which of the following is NOT a feature of a high-quality multiple-choice (MC) item?	- measures knowledge consistently, accurately - measures knowledge supposed to be measured - includes at least four response options
5. Three types of selected-response (SR) items are alternative response, matching, and multiple-choice.	- True - False
6. Multiple-choice tests would provide more reliable scores than true-false tests because	- more consistency of scores is obtained - scoring is more objective - the effect of guessing is reduced
7. A halo effect happens when extraneous factors (e.g., student behaviors) influence the teacher's scoring of constructed-response (CR) items.	- True - False

Table B3

Items in CBA with the Selected-Response Test 3

Item	Response options
1. A multiple-choice item with a difficulty of 0.75 and discrimination of 0.38 is easier and discriminates better between higher and lower achievers compared to an item with a difficulty of 0.25 and discrimination of 0.13.	- True - False
2. Which of the following actions in preparing a portfolio MOST strongly promotes student academic growth?	- students choose which samples to include - students reflect on their own work - students collect feedback from teachers
3. Grading-wise, students who bring cookies should NOT receive bonus marks in order to retain fairness in grading practices.	- True - False
4. In absolute (criterion-referenced) grading, the same proportion of students receive honors from year to year.	- True - False
5. An item analysis of a multiple-choice test can provide information to the teacher regarding	- misconceptions on specific topics. - time to complete the test. - which students guessed.
6. In relative (norm-referenced) grading, it is hard to maintain the same standards across years.	- True - False
7. Which of the following assessments should NOT be used by teachers as part of the grade calculation process?	- formative tasks - portfolio assessments - performance assessments
8. In Alberta, teachers are required to use norm-referenced grading.	- True - False

Table B4

Item Parameters and Test Reliability for CBA with Selected-Response Tests

	SR test 1		SR test 2		SR test 3	
	b	*a*	*b*	*a*	*b*	*a*
Item 1	0.95	0.09	0.91	0.16	0.62	0.41
Item 2	0.95	0.09	0.77	0.38	0.79	0.34
Item 3	0.92	0.09	0.84	0.14	0.96	0.07
Item 4	0.82	0.18	0.75	0.12	0.83	0.34
Item 5	0.89	0.21	0.90	0.09	0.95	0.10
Item 6	0.61	0.55	0.82	0.28	0.57	0.17
Item 7	0.91	0.18	1	0	0.83	0.27
Item 8	0.83	0.34	—	—	0.91	0.17

Note. SR: selected-response; *b*: item difficulty parameter; *a*: item discrimination parameter. Reliability coefficients are 0.38, 0.31, and 0.51, respectively.

Appendix C

Self-Assessment in CBA

Once students finish CBA, they are asked to self-assess their performance with the question "*Thank you for reviewing some topics with me. How would you rate your own performance on these questions? Inadequate, Moderate, or Excellent"* Depending on their response category, they are provided final support.

Self-assessment category	Support
Inadequate	These can be tricky topics, but they are important not only for your final exam but for your IPT practicum. Please make a plan to study these topics more. You can review your notes or contact a TA for support.
Moderate	I think you are well on your way to understanding these topics, but it is a good idea to have a plan. You can review your notes or contact a TA for support.
Excellent	Great! You are on solid footing but remember that these questions were lower level than what you'll see on the final exam so be sure to continue to review and focus on applying the information.

Appendix D

Survey Questions

In the final part of CBA, students are invited to participate in a survey with the question "*Do you have five more min? I would love to hear your thoughts on working with a Chatbot. Click the link and let me know what you think! Bye!"*

Table D1

Likert Scale Survey Items

Likert scale survey items Strongly Disagree, Disagree, Agree, Strongly Agree, Not sure
This section includes a set of questions focusing on experience with the chatbot.
I found the feedback in the chatbot helpful.
The feedback helped me stay motivated.
I found the summary answer in the chatbot helpful.
The summary answer helped me improve my existing understanding of the concept.
I felt comfortable when interacting with the chatbot.
I was engaged during the assessment.
I put enough effort to answer each question.
The conversations in the chatbot helped me stay focused.
I found taking an assessment with the chatbot straightforward.
This section includes a set of questions focusing on feelings towards chatbots.
I prefer to take a practice exam with a chatbot compared to an online quiz.
I would perform better in a chatbot than an online quiz.
Chatbot would provide a more accurate representation of my performance than an online quiz.

Table D2

Demographic and Background Items

Item	Options
What is your age?	
What is your gender?	Female Male Non-binary Prefer not to say
Currently, what year of university are you in?	1st year 2nd year 3rd year 4th year 5th + year
Have you ever had experience with a chatbot before?	Yes No Not sure
How would you rate your current technological skills?	Beginner Intermediate Expert
How would you rate your confidence to answer questions related to the Educational Assessment subject?	I am not confident I am somewhat confident I am confident
Have you faced any problems when using the chatbot?	No Yes, some minor problems Yes, some major problems